BEHAVIORAL AND BRAIN SCIENCES (2009) **32**, 121–182
Printed in the United States of America
doi:10.1017/S0140525X09000545

How we know our own minds: The relationship between mindreading and metacognition

Peter Carruthers
Department of Philosophy, University of Maryland, College Park, MD 20742
pcarruth@umd.edu
http://www.philosophy.umd.edu/Faculty/pcarruthers/

Abstract: Four different accounts of the relationship between third-person mindreading and first-person metacognition are compared and evaluated. While three of them endorse the existence of introspection for propositional attitudes, the fourth (defended here) claims that our knowledge of our own attitudes results from turning our mindreading capacities upon ourselves. Section 1 of this target article introduces the four accounts. Section 2 develops the "mindreading is prior" model in more detail, showing how it predicts introspection for perceptual and quasi-perceptual (e.g., imagistic) mental events while claiming that metacognitive access to our own attitudes always results from swift unconscious self-interpretation. This section also considers the model's relationship to the expression of attitudes in speech. Section 3 argues that the commonsense belief in the existence of introspection should be given no weight. Section 4 argues briefly that data from childhood development are of no help in resolving this debate. Section 5 considers the evolutionary claims to which the different accounts are committed, and argues that the three introspective views make predictions that are not borne out by the data. Section 6 examines the extensive evidence that people often confabulate when self-attributing attitudes. Section 7 considers "two systems" accounts of human thinking and reasoning, arguing that although there are introspectable *events* within System 2, there are no introspectable *attitudes*. Section 8 examines alleged evidence of "unsymbolized thinking". Section 9 considers the claim that schizophrenia exhibits a dissociation between mindreading and metacognition. Finally, section 10 evaluates the claim that autism presents a dissociation in the opposite direction, of metacognition without mindreading.

Keywords: Autism; confabulation; conscious thought; introspection; metacognition; mindreading; schizophrenia; self-interpretation; self-monitoring; self-knowledge

1. Introduction

Human beings are inveterate mindreaders. We routinely (and for the most part unconsciously) represent the mental states to the people around us (thus employing *metarepresentations* – representations of representational states). We attribute to them perceptions, feelings, goals, intentions, knowledge, and beliefs, and we form our expectations accordingly. While it isn't the case that *all* forms of social interaction require mindreading (many, for example, follow well-rehearsed "scripts" such as the procedures to be adopted when boarding a bus or entering a restaurant), it is quite certain that without it, human social life would be very different indeed. But human *mental* life, too, is richly metarepresentational, containing frequent attributions of mental states to *ourselves*. This sort of first-person metarepresentation is generally referred to as "metacognition." The present target article is about the cognitive basis (or bases) of our dual capacities for mindreading and for metacognition, and the relationships between them. For reasons that emerge in section 2, however, our main focus will be on *propositional attitude* mindreading and metacognition (involving attributions of beliefs, judgments, intentions, decisions, and the like) rather than on our capacities for attributing mental states more generally.

At least four different accounts of the relationships that obtain between mindreading and metacognition can be distinguished. Three of them maintain that our access to our own minds is quite different in *kind* from our access to the minds of other people (because they involve a form of introspection), whereas the fourth (which will be defended here) denies this. The present section provides a brief explanation of each, before making some further introductory comments.

1.1. Model 1: Two mechanisms

One possibility is that mindreading and metacognition are two independent capacities, realized in distinct cognitive mechanisms. Nichols and Stich (2003) have elaborated

PETER CARRUTHERS is Professor of Philosophy at the University of Maryland, College Park. He is the author or co-author of eleven books and co-editor of seven, and has published around a hundred articles and reviews. Most of his recent work has concerned reductive explanations of phenomenal consciousness, the involvement of language in thought, mental modularity, and the character of self-knowledge. His most recent book is *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*, published in 2006 by Oxford University Press.

© 2009 Cambridge University Press 0140-525X/09 $40.00

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

and defended this view. Their model of the mindreading system is an eclectic one, involving both simulation-like aspects and information-rich components (both theory-like and modular). There are postulated mechanisms for detecting the perceptual states of other people, for detecting the desires of other people, and for detecting the beliefs of other people where they differ from our own. A "Possible Worlds Box," or hypothetical reasoning system, is used to construct a representation of the world as seen by the other person (containing as suppositions the beliefs and goals attributed to the other), and then the subject's own inferential and planning mechanisms are used to figure out what else the target might believe, or to work out what the target might do. (Crucially, and in distinction from most other forms of simulation theory, this stage isn't supposed to involve introspection of one's own states.) While most of the basic components are held to be innate, there is said to be much work left for learning to do in the course of childhood development.

When Nichols and Stich (2003) then turn to provide an account of self-awareness, they defend the view that there are two (or more) distinct self-monitoring mechanisms. There is at least one such mechanism for monitoring and providing self-knowledge of our own experiential states, and one (at least) for monitoring and providing self-knowledge of our own propositional attitudes. These mechanisms are held to be distinct from one another, and also from the mindreading system that deals with the mental states of other people. They are also held to be innate and to emerge under maturational constraints early in infancy.

An account of this sort predicts a double dissociation between mindreading and metacognitive capacities. Since these are held to be realized in two (or more) independent mechanisms, there should exist cases where each is damaged or interfered with in the absence of damage or interference occurring to the other (Sternberg 2001). So there should be cases of people who can attribute mental states to others successfully but who have difficulty in attributing mental states to themselves, as well as instances of people who maintain reliable access to their own mental states while losing their capacity to attribute such states to other people. Nichols and Stich (2003) argue that people with passivity-symptom schizophrenia fit the first profile, whereas people with autism fit the second, thus confirming their account. These arguments are discussed and evaluated further on.

1.2. Model 2: One mechanism, two modes of access

A second account maintains that there is just a single metarepresentational faculty, but one that has two distinct kinds of access to the mental states with which it deals, using distinct informational channels. This single faculty has both a perception-based mode, used when interpreting other people, and an introspective mode, used when accessing and representing one's own mental states. Although it is unclear whether such a view has ever been defended explicitly in print, it is implicit in Carruthers (1996a), and it is often suggested in conversation, especially among those who endorse a "modular" account of the mindreading faculty. Moreover, both Frith and Happé (1999) and Happé (2003) are quite naturally interpreted in this way (although they might also be read as endorsing model 4).

This account has one significant advantage over the "two independent mechanisms" proposal just considered: It provides a smooth and natural explanation of the fact that self-knowledge and other-knowledge utilize the same conceptual resources. This will be because the very same concepts and/or the very same body of "core knowledge" of mental states are housed in one and the same metarepresentational faculty, albeit a faculty that has input connections deriving not only from the mental states of other people (indirectly, via perception) but also from oneself (more directly, via introspection).

This sort of single-mechanism account makes slightly different predictions regarding the expected dissociations. Like model 1, it entails that there should be cases in which self-knowledge is compromised (because the introspective inputs to the metarepresentational faculty have been disrupted), whereas other-knowledge is intact (because the faculty itself remains undamaged and still has access to perceptual input). And it predicts that there should be cases where *both* self-knowledge *and* other-knowledge are compromised, by virtue of damage to the metarepresentational faculty itself. (Frith and Happé [1999] can quite naturally be interpreted as arguing that people with autism fit this profile.) But there should be *no* cases where other-knowledge is damaged while self-knowledge is left intact, except by virtue of massive multi-modal perceptual failure.[1] These predictions, too, are examined in due course.

1.3. Model 3: Metacognition is prior

A third view maintains that metacognition is prior to mindreading, in such a way that the attribution of mental states to others depends upon our introspective access to our own mental states, together with processes of inference and simulation of various sorts. Goldman (1993; 2006), among others, have proposed and defended accounts of this kind. They also lie behind much of the excitement surrounding the discovery of so-called mirror neurons (Gallese & Goldman 1998; Gallese et al. 1996; Rizzolatti et al. 1996). For it is by virtue of awareness of our own action-tendencies, caused by observing the actions of others, that we are supposed to gain our initial social understanding.

Goldman's account of our introspective abilities has evolved over the years. In his 1993 target article, he thought that our access to our own propositional attitudes was mediated via awareness of the phenomenal feelings that are distinctive of them. This view came in for heavy criticism, however (Carruthers 1996c; Nichols & Stich 2003), and he now maintains that introspection uses an innate code in the language of thought, whose basic elements are caused by the various mental state types, responding to features of their neural realization (Goldman 2006). But the account of mindreading remains essentially the same: One adopts, in imagination, the perspective of a target subject, reasons on one's own behalf within the scope of that imagination (hence simulating the reasoning processes of the other), and then introspects the resulting mental state of belief or decision, before attributing such a state to the agent in question.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Model 3 makes predictions similar to those of model 2, but with an opposite valence. Both accounts agree that there should be cases in which both mindreading and metacognition are damaged. (In the case of Goldman's model, this will occur whenever the introspective capacity is disrupted, since mindreading is held to be grounded in introspective access to one's own mind.) But instead of predicting that there should be cases where metacognition is poor while mindreading is normal, as did model 2, the present account predicts the opposite: that there should be cases where metacognition is normal while mindreading is damaged. This would happen whenever the simulative abilities utilized in mindreading are disrupted. Following Nichols and Stich (2003), Goldman (2006) argues that people with autism fit this profile.

1.4. Model 4: Mindreading is prior

A fourth view, in contrast, claims the reverse of the third: Instead of mindreading being grounded in metacognition, it maintains that metacognition is merely the result of us turning our mindreading capacities upon ourselves. A variety of different versions of such an account have been proposed (Carruthers 2006; Gazzaniga 1995; 2000; Gopnik 1993; Wegner 2002; Wilson 2002; some differences among these authors will emerge as we proceed).[2] The purpose of the present target article is to explain, elaborate, and defend the most plausible variant of this final sort of view. Section 2 will embark on that task.

This fourth account entails that there should be *no* dissociations between mindreading and metacognition. This is because there is just a single faculty involved in both forms of activity, using essentially the same inputs, which are all perceptual or quasi-perceptual in character (including visual imagery and "inner speech"; see sect. 2). However, the account also predicts that it should be possible to induce subjects to *confabulate* attributions of mental states to themselves by manipulating perceptual and behavioral cues in such a way as to provide misleading input to the self-interpretation process (just as subjects can be misled in their interpretation of others). Likewise, the account predicts that there should be no such thing as awareness of one's own propositional attitudes independently of any perceptually accessible cues that could provide a basis for self-interpretation. The accuracy of these predictions will be discussed and evaluated in due course. Note that the "mindreading is prior" account is the only one of the four to make such predictions.

Notice that each of the first three accounts just described endorses the existence of some variety or other of *introspection*, understood broadly to encompass any reliable method for forming beliefs about one's own mental states that *is not* self-interpretative and that differs in *kind* from the ways in which we form beliefs about the mental states of other people. (It should be emphasized that the term "introspection" is used in this broad, negatively defined, sense throughout this target article. Many different specific views are thereby included.) Notice that to say that an introspective process is not self-interpretative doesn't mean that it isn't *inferential*. On the contrary, those who take seriously the analogy between introspection and external perception, and who think that the former is realized in a self-monitoring mechanism of some sort, are apt to think that it achieves its output by effecting computations on the data that it receives as input (just as does vision, for example). But these inferences will presumably rely on general principles, such as (in the case of vision) that light shines from above, or that moving objects are locally rigid. For present purposes, an *interpretative* process, in contrast, is one that accesses information about the subject's current circumstances, or the subject's current or recent behavior, as well as any other information about the subject's current or recent mental life. For this is the sort of information that we must rely on when attributing mental states to other people.

In contrast with the first three accounts, proponents of view 4, who maintain that metacognition results from us turning our mindreading abilities upon ourselves, must deny the existence of introspection (at least for a significant class of mental states; see sect. 2). So also at stake in this target article is the commonsense view that we have introspective access to our own minds (or at least to certain aspects of them).

2. Elaborating the "mindreading is prior" model

As we noted earlier, a number of different versions of the "mindreading is prior" view have been proposed. These come in different strengths. At one extreme is Gopnik (1993). In her target article on this topic, she urged that the attribution of *all* mental states to oneself (with the exception, perhaps, of what she described as some sort of "Cartesian buzz") is equally theory-based, and equally interpretational. But this strong view has come in for heavy criticism. For as Nichols and Stich (2003) and Goldman (2006) both point out, I seem to be able to know what I am currently thinking and planning even though I am sitting quiet and motionless (in which case there will be no behavior available for the mindreading system to interpret). How is this possible, the critics ask, unless we have access to our own mental states that isn't interpretative, but is rather introspective?

At the other extreme lie Wegner (2002) and Wilson (2002), who are often interpreted as proponents of a "mindreading is prior" account. Each makes a powerful case that we *often* attribute propositional attitudes to ourselves via self-interpretation (and often false and confabulated interpretation, at that). But both seem to allow that we *also* have access to *some* of our attitudes that is introspective in character. For each allows that we undergo conscious as well as unconscious thoughts, and that the former can provide part of the evidence base for self-attributing the latter. I argue in section 7 that they have been misled, however, and that they have run together the sensory accompaniments of attitudes – such as inner speech and visual imagery (to which we do have introspective access, I allow) – with the attitudes themselves.

In contrast with the preceding accounts, the position to be defended in the present target article is as follows. There is just a single metarepresentational faculty, which probably evolved in the first instance for purposes of mindreading (or so I shall argue in sect. 5). In order to do its work, it needs to have access to perceptions of the environment. For if it is to interpret the actions of others, it plainly requires access to perceptual representations of those actions.[3] Indeed, I suggest that, like

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

most other conceptual systems, the mindreading system can receive as input any sensory or quasi-sensory (e.g., imagistic or somatosensory) state that gets "globally broadcast" to all judgment-forming, memory-forming, desire-forming, and decision-making systems. (For evidence supporting a global broadcasting cognitive architecture, see Baars 1988; 1997; 2002; 2003; Baars et al. 2003; Dehaene & Naccache 2001; Dehaene et al. 2001; 2003; Kreiman et al. 2003.)

By virtue of receiving globally broadcast perceptual states as input, the mindreading system should be capable of self-attributing those percepts in an "encapsulated" way, without requiring any other input. Receiving as input a visual representation of a man bending over, for example, it should be capable of forming the judgment, "I am seeing a man bending over." (At least, this should be possible provided the visual state in question has been partially conceptualized by other mental faculties, coming to the mindreading system with the concepts *man* and *bending over* already attached. I return to discuss the significance of this point shortly.) This is the way in which introspection of perceptual, imagistic, and somatosensory mental events is achieved, I suggest. Given that the mindreading faculty possesses the concepts *sight*, *hearing*, and so forth (together with a concept of self), it should be able to activate and deploy those concepts in the presence of the appropriate sort of perceptual input on a recognitional or quasi-recognitional basis (Carruthers 2000). Because no appeals to the subject's own behavior or circumstances need to be made in the course of making these judgments, the upshot will qualify as a form of introspection, in the broad sense being used here.

Let me stress, however, that what is being offered here is an account of *introspection* for perceptual states, not an account of experiential, or "phenomenal," *consciousness*. (And although I sometimes use the language of "consciousness" in this target article, this should always be understood to mean *access* consciousness rather than *phenomenal* consciousness; see Block [1995] for the distinction.) Although global broadcasting is often put forward as a theory of phenomenal consciousness (Baars 1988; 1997), that isn't how it is being used in the present context. Rather, it forms part of an account of how we come to have knowledge of our own perceptual and quasi-perceptual states. Whether global broadcasting provides a sufficient explanation of the "feely" qualities of phenomenal consciousness is another matter entirely. And although I myself have defended a higher-order account of phenomenal consciousness, according to which it is the availability of globally broadcast states to the mindreading faculty that is responsible for their phenomenally conscious status (Carruthers 2000), I don't mean to rely on that here, either. Indeed, I intend the discussion in this target article to be neutral among proposed explanations of phenomenal consciousness.

Although the mindreading system has access to perceptual states, the proposal is that it lacks any access to the outputs of the belief-forming and decision-making mechanisms that feed off those states. Hence, self-attributions of propositional attitude events like judging and deciding are always the result of a swift (and unconscious) process of self-interpretation. However, it isn't just the subject's overt behavior and physical circumstances that provide the basis for the interpretation. Data about perceptions, visual and auditory imagery (including sentences rehearsed in "inner speech"), patterns of attention, and emotional feelings can all be grist for the self-interpretative mill.

Such an account can plainly avoid the difficulties that beset Gopnik (1993). For consider someone sitting quietly in his living room, who has just become aware of deciding to walk to his study to get a particular book from the shelf (Goldman 2006, p. 230). His mindreading system has access to a variety of forms of evidence in addition to overt behavior (which in this case is lacking). The agent might, for example, have verbalized or partially verbalized his intention, in "inner speech." And then, since inner speech utilizes the same perceptual systems that are involved in the hearing of speech (Paulescu et al. 1993; Shergill et al. 2002), this will be available as input to the mindreading system. Or he might have formed a visual or proprioceptive image of himself selecting that particular book, which will be similarly available (Kosslyn 1994). Or the context provided by his prior verbalized thoughts and visual images, together with a shift in his attention towards the door, might make it natural to interpret himself as having decided to walk to his study to collect that particular book.

Notice that allowing the mindreading system to have access to visual imagery, proprioceptive data, and emotional feelings is pretty much mandatory once we buy into a global broadcasting architecture, even though such events will presumably play little or no role in third-person mental-state attribution. For perceptual and quasi-perceptual states of all kinds are capable of being globally broadcast when attended to, and will thus become available to any conceptual system that looks to such broadcasts for its input. But the upshot is to blur the boundary somewhat between the "mindreading is prior" account and model 2 ("one mechanism, two modes of access"). For we now have to concede that the mindreading system does have available to it information when attributing mental states to the self that it never has access to when attributing mental states to others. For unless subjects choose to tell me, I never have access to what they are imagining or feeling; and certainly I never have the sort of direct access that my mindreading system has to my own visual images and bodily feelings.

Despite this "blurring of boundaries," there remains good reason to insist on the distinctness of our account from model 2. This is because the latter is committed to the claim that the metarepresentational faculty has introspective, non-interpretative access to mental states of all types, including propositional attitudes as well as sensory experiences. The account being proposed here, in contrast, maintains that our access to our own propositional attitudes is *always* interpretative, while conceding that the evidence base for self-interpretation is somewhat wider than we normally have available when interpreting other people.

One final point needs to be emphasized: As the example of seeing a man bending over should make clear, the thesis that judgments aren't introspectable requires important qualification. In particular, it should be restricted to judgments that aren't perceptual judgments. According to Kosslyn (1994) and others, the initial outputs of the visual system interact with a variety of conceptual systems that deploy and manipulate perceptual templates, attempting to achieve a "best match" with the incoming

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

data. When this is accomplished, the result is globally broadcast as part of the perceptual state itself. Hence, we see an object *as* a man or *as* bending over. Because this event is apt to give rise immediately to a stored belief, it qualifies as a (perceptual) judgment. But because it will also be received as input by the mindreading system (by virtue of being globally broadcast), it will also be introspectable. In the discussion that follows, therefore, whenever I speak of "judgments," I should be understood to mean "*non-perceptual* judgments," such as the judgment that 17 is a prime number or that polar bears are endangered.[4]

2.1. Mindreading and speech

If we lack introspective access to our own propositional attitudes, then how is it that we can report on those attitudes, swiftly and unhesitatingly, in the absence of anything that could plausibly be seen as input to a process of self-interpretation? If someone asks me for the date on which I think the Battle of Hastings took place, for example, I can reply immediately, "1066, I believe." But on what basis could I interpret myself as possessing such a belief? I can recall no Battle-of-Hastings-related behavior; and there need have been nothing relevant of an imagistic sort passing through my mind at the time, either.

There is surely no reason to think, however, that the verbal expression of a belief requires prior metacognitive access to it. Rather, one's executive systems will conduct a search of memory, retrieving an appropriate first-order content which can then, in collaboration with the language faculty, be formulated into speech. And then attaching the phrase, "I think that . . ." or "I believe that . . ." to the first-order sentence in question is a trivial matter (Evans 1982), and is often a mere manner of speech or a matter of politeness (so as not to appear too confident or too definite). It certainly needn't require that subjects should first formulate a metacognitive judgment to the effect that they believe the content in question. Hence, it may be that the first metacognitive access subjects have to the fact that they have a particular belief is via its verbal expression (whether overtly or in inner speech). And such speech, like all speech, will need to be interpreted to extract its significance.

General considerations of cognitive engineering support such a view. For we already know that executive systems would need to have access to stored information, and that they would have been honed by evolution to conduct efficient searches for the information required to solve each type of practical task in hand. Moreover, this capacity would surely have been of ancient evolutionary provenance, long pre-dating the emergence of language and mindreading. Nor does it qualify as a form of introspection, since it isn't metarepresentational in character. When the mindreading system was added in the course of human evolution, therefore, there would have been no *need* for it to be built with its own capacities to conduct searches of all memory; and on the contrary, since all data-mining is computationally expensive, this would have come at significant additional cost. And while there is every reason to think that capacities for language and for mindreading would have coevolved (Gomez 1998; Origgi & Sperber 2000), there isn't any reason to think that the language faculty can only produce an output when provided with a metacognitive content as input, either issued by the mindreading faculty or by a separate faculty of introspection.

Many cognitive scientists think that the speech-production process begins with a thought-to-be-expressed (Levelt 1989). I myself believe that this is an exaggeration (Carruthers 2006). Speech is an action, and like other actions can be undertaken for a variety of purposes (the expression of belief being only one of them). Hence, any utterance in the indicative mood needs to be interpreted to determine whether it is made ironically, or in jest, or as a mere supposition; or whether it is, indeed, expressive of belief. However, I know of *no* theorist who thinks that speech needs to begin from a *metacognitive* representation of the thought to be expressed. So even utterances that do express a corresponding belief don't qualify as a form of introspection, since no metarepresentational *thought* occurs until one's own words are heard and interpreted.

Similar points hold in respect to the verbal expression of desire. No doubt we often give voice to our desires, having first envisaged the thing or circumstance in question and monitored and interpreted our affective responses, in the manner proposed by Damasio (1994; 2003). (This is, of course, fully consistent with a "mindreading is prior" account.) But often our current desires can recruit appropriate speech actions in their own service, with use of the terminology of "want" or "desire" being just one possible means among many. Thus, the two-year-old child who says, "I want juice," is unlikely to have *first* formulated a metacognitive thought. Rather, desiring juice, the child is seeking ways to achieve that goal. And for these purposes a number of different speech actions might be equally effective, including, "Give me juice," "Juice, please," and so on. If she chooses to say, "I want juice," then she does make an assertion with a metacognitive content, and hence (if she understands the concept of wanting) she will *subsequently* come to entertain a metacognitive thought. But there is no reason to think that her utterance must begin with such a thought, any more than does the utterance of someone who answers the question, "Is it the case that P?" by saying, "Yes, I believe that P."

It might be objected that even if we sometimes learn of our own beliefs and desires by first becoming aware of their formulation into speech (whether inner or outer), this still gives us reliable, non-interpretative access to them. Hence this can still count as a form of introspection. But this appearance of immediacy is illusory. All speech – whether the speech of oneself or someone else – needs to be interpreted before it can be understood. Unless we beg the point at issue and assume that subjects have direct introspective access to their own articulatory intentions, the language-comprehension system will need to get to work on the utterance in the normal way, figuring out its meaning in light of the utterance's linguistic properties (lexical meanings, syntax, etc.) together with knowledge of context. And even if, as is likely, the result of this process (the content of the utterance) is attached to the existing representation of the sound of the utterance and globally broadcast to all conceptual systems, including the mindreading faculty, the latter will still only have interpretative access to the underlying beliefs or goals that initiated the utterance.

But how is it, then, that our own utterances are not ambiguous to us, in the way that the utterances of other

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

people often are? If I find myself thinking, "I shall walk to the bank," then I don't need to wonder which sort of bank is in question (a river bank, or a place where one gets money). And this fact might be taken to indicate that I must have introspective access to my intentions. However, there will generally be cues available to disambiguate our own utterances, which wouldn't be available to help interpret the similar utterances of another. For example, just prior to the utterance I might have formed a visual image of my local bank, or I might have activated a memory image of an empty wallet. But even when no such cues are available, there remains a further factor that will serve to disambiguate my own utterances, but which won't always help with the utterances of others. This is the relative *accessibility* of the concepts involved, which is a pervasive feature of speech comprehension generally (Sperber & Wilson 1995). Because the goals that initiated the utterance, "I shall walk to the bank," would almost certainly have included an activation of one or other specific concept *bank*, this will ensure the increased accessibility of that concept to the comprehension system when the utterance is processed and interpreted.

I conclude, therefore, that while subjects can often express their beliefs in speech, and can hence acquire more-or-less reliable information about what they believe, this gives us no reason to think that introspection for propositional attitudes exists.

3. The introspective intuition

There is no doubt that the denial of introspection for propositional attitudes, entailed by the "mindreading is prior" view, is hugely counterintuitive to most people. Almost every philosopher who has ever written on the subject, for example – from Descartes (1637), Locke (1690), and Kant (1781), through to Searle (1992), Shoemaker (1996), and Goldman (2006) – has believed that many (at least) of our own judgments and decisions are immediately available to us, known in a way that is quite different from our knowledge of the judgments and decisions of other people. We are (pre-theoretically) strongly inclined to think that we don't need to *interpret* ourselves in order to know what we are judging or deciding (or that we don't need to do so all of the time, at least; many of us now have enough knowledge of cognitive science to concede that such events can also occur unconsciously). Rather, such events are often (somehow) directly available to consciousness. Since it is generally thought to be a good thing to preserve intuitions ceteris paribus, this might be taken to create a presumption in favor of one of the three alternative accounts that we considered at the outset. The strategy of this section is to draw the teeth from this argument by showing, first, that the intuition underlying it is unwarranted, and then by using reverse engineering to explain why (from the perspective of a "mindreading is prior" account) it nevertheless makes good sense that such a folk-intuition should exist.

3.1. The subjective experience of introspective access isn't evidence of introspection

The thesis expressed in this subsection's title is clearly demonstrated by research with commissurotomy ("split-brain") subjects, conducted over many years by Gazzaniga (1995; 2000) and colleagues. In one famous case (representative of many others of similar import), Gazzaniga (1995) describes how different stimuli were presented to the two half-brains of a split-brain patient simultaneously. The patient fixated his eyes on a point straight ahead, while two cards were flashed up, one positioned to the left of fixation (which would be available only to the right brain) and one to the right of fixation (which would be available only to the left brain). When the instruction, "Walk!" was flashed to the right brain, the subject got up and began to walk out of the testing van. (The right brain of this subject was capable of some limited understanding of language, but had no production abilities.) When asked why, he (the left brain, which controlled speech-production as well as housing a mindreading system) replied, "I'm going to get a Coke from the house." This attribution of a current intention to himself is plainly confabulated, but delivered with all of the confidence and seeming introspective obviousness as normal.

It is important to note that although commissurotomy patients can often have a good understanding of their surgery and its effects, they never say things like, "I'm probably choosing this because I have a split brain and the information went to the right, non-verbal, hemisphere" (Gazzaniga 1995). On the contrary, they make their confabulated reports smoothly and unhesitatingly, and their (i.e., their left brain's) sense of self seems quite unchanged following the operation. Even reminders of their surgery during testing have no effect. On a number of occasions testing was paused and the experimenter said something like, "Joe, as you know, you have had this operation that sometimes will make it difficult for you to say what we show you over here to the left of fixation. You may find that your left hand points to things for that reason, OK?" Joe assents, but then on the very next series he is back to showing the interpreter effect once again (Gazzaniga, personal communication). If patients were aware of interpreting rather than introspecting, then one would expect that a reminder of the effects of commissurotomy would enrich the hypothesis pool, and would sometimes lead them to attribute some of their own behavior to that. But it doesn't do so.

Of course it doesn't follow from the extensive commissurotomy data that normal human subjects never have privileged, non-interpretative, access to their own judgments and decisions, as Goldman (2006) points out. (And for this reason the defense of a "mindreading is prior" account that is mounted by Gazzaniga [1998] strikes many people as massively under-supported. One way of viewing the present target article is that it is an attempt to rectify that deficiency.) Gazzaniga's data were collected from patients who had undergone serious brain damage (a severed corpus collosum). Hence, it may be that in normal brains the mindreading system does have immediate access to the agent's judgments and intentions. The split-brain data force us to recognize that *sometimes* people's access to their own judgments and intentions can be interpretative (much like their access to the judgments and intentions of other people), requiring us at least to accept what Goldman (2006) calls a "dual method" theory of our access to our own thoughts. But one could believe (as Goldman does) that introspection is the normal, default, method for acquiring knowledge

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

of our own propositional attitudes, and that we only revert to self-interpretation as a back-up, when introspection isn't available.

The split-brain data show decisively that we don't have any *introspective*, *subjectively accessible*, warrant for believing that we ever have introspective access to our own judgments and decisions, however. This is because patients report plainly confabulated explanations with all of the same sense of obviousness and immediacy as normal people. And if normal people were to rely upon subjectively accessible cues to identify cases of introspection, then commissurotomy patients should be able to use the absence of such cues to alert them to the interpretative status of their reports. The best explanation is therefore that subjects themselves can't tell when they are introspecting and when they are interpreting or confabulating. So for all we know, it may be that our access to our own judgments and decisions is *always* interpretative, and that we *never* have introspective access to them. Now philosophers will note, of course, that given so-called reliabilist conceptions of knowledge and justification, we might count as knowing, and as justified in believing in, the existence of introspection, despite our inability to discriminate cases of introspection from cases of confabulation. This will be so provided that introspection really does exist and is common, and provided that our belief in it is reliably caused by the fact that we do often introspect, and is caused in the right sort of way. My point here, however, is that our inability to discriminate shows that we have no *subjectively accessible* reason to believe in the existence of introspection. So anyone who is wondering whether or not introspection is real should realize that they have no reason they can offer for thinking that it is, in advance of examining the evidence.

3.2. The mindreading system's model of its own access to the mind

The intuition that there is introspection for propositional attitudes is therefore unwarranted. But in addition, we can explain why we should have such an intuition in the first place, even if (as I am suggesting) it turns out to be false. This is because the mindreading system's operations will be greatly simplified, but without any significant loss of reliability (and perhaps with some gain), if its model of its own access to the mind is an introspective (non-interpretative) one. We should then predict that just such a model would be arrived at, whether by natural selection or through individual learning. This argument is laid out in some detail in Carruthers (2008a). In consequence, this section is brief.[5]

In order to be effective, the mindreading system needs to contain some sort of model of the way that minds, in general, work. Such a model should include an account of the access that agents have to their own mental states. And here there are essentially two choices. The mindreading system can either represent agents as interpreters of themselves, or it can picture them as having direct introspective access to their own mental states. The former would complicate the mindreading system's computations, and would mandate consideration of a wider range of evidence, taking into account the possibility of misinterpretation. But there is unlikely to be any compensating gain in reliability. One reason for this is that people are, probably, excellent interpreters of themselves. (We know that they are remarkably good interpreters of others.) Hence, in normal circumstances instances of confabulation will be rare, and thus any errors introduced by a belief in introspection will be few. A second reason is that self-attributions of mental states, even if initially confabulated, are likely to be self-fulfilling. This is because agents will feel obliged to act in ways that are consistent with the mental states that they have attributed to themselves. And a third reason is that any expansion in the computational complexity of a system will introduce additional sources of error (as well as imposing a cost in terms of speed of processing, of course), as will any increase in the types of evidence that need to be sought. It is now a familiar point in cognitive science, not only that simple (but invalid) heuristics can prove remarkably reliable in practice, but that they can often out-compete fancier computational processes once the costs imposed by computational errors, as well as missing or misleading information, are factored in (Gigerenzer et al. 1999).[6] What we should predict, therefore, is that the mindreading system should model the mind as having introspective access to itself. And then that very same model will render agents blind to the fact (if it is a fact) that their mode of access to their own mental states is actually an interpretative one.

I conclude that the playing field is now leveled between the competing theories, in the sense that there is no initial presumption against model 4. And given a level playing field, we should prefer the simplest theory ceteris paribus. This means that the "mindreading is prior" account should now be our default option, because it postulates just a single mechanism with a single mode of access to its domain, whereas the other accounts postulate greater complexity.

4. The data from development

Gopnik (1993) bases much of her case for a "mindreading is prior" account on developmental evidence, claiming that there is a parallelism between children's performance in mindreading tasks and matched metacognitive tasks (see also, Gopnik & Meltzoff 1994). This claim has held up well over the years. In an extensive meta-analysis of hundreds of experiments, Wellman et al. (2001) are able to find no evidence of any self/other asymmetry in development. Taken at face value, these data count strongly against both a "two independent mechanisms" account and a "metacognition is prior" view, each of which predicts that metacognitive competence should emerge in development in advance of mindreading.

What most parties in these debates have overlooked, however, is the existence of the remaining alternative to a "mindreading is prior" account, namely the "one mechanism, two modes of access" view. For this, too, predicts that development in the domains of both self- and other-understanding should proceed in parallel. Like the "mindreading is prior" view, this account claims that there is just a single mechanism or body of core knowledge underlying both mindreading and metacognitive competence. Hence, one would expect children's capacities in both domains to emerge at about the same time. What this means is that developmental evidence is inherently incapable of

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

discriminating between views that endorse, and those that deny, the existence of introspective access to propositional attitudes.

There is another, equally important, reason why developmental evidence is of no use to us in this inquiry, however. This is that all parties in the debate over the existence of introspection for attitudes have shared a traditional and widely accepted understanding of the developmental timetable for mindreading competence (Goldman 2006; Gopnik 1993; Nichols & Stich 2003). This was thought to proceed through well-defined stages over the first four or five years of life, with competence in false-belief reasoning not emerging until after the age of four (Wellman 1990). Yet there have always been those who have maintained that an underlying competence with false-belief might be present much earlier, but masked by young children's difficulties in executive functioning (Fodor 1992; Leslie & Polizzi 1998). Indeed, Birch and Bloom (2004; 2007) refer to the latter as "the curse of knowledge," pointing out that adults, too, can often have difficulty in allowing for the false beliefs of another. And this general perspective has now received dramatic confirmation through the use of nonverbal looking-time and expectation measures. These show competence with false-belief understanding and other allegedly late-emerging aspects of mindreading capacity at around 15 or 24 months, *long* before this had traditionally been thought possible (Bosco et al. 2006; Onishi & Baillargeon 2005; Onishi et al. 2007; Song & Baillargeon, forthcoming; Song et al., forthcoming; Southgate et al. 2007; Surian et al. 2007). But no one has, as yet, been able to develop nonverbal measures of metacognitive understanding in infants for purposes of comparison.

Of course there is much here that needs to be explained. In particular, if metarepresentational competence is present in the second year of life, we want to know why it takes two or more additional years for that competence to manifest itself in verbally based tasks. But this isn't a question for us. Our focus is on adjudicating between accounts that endorse the existence of introspection and those that deny it. And for these purposes it is plain that we need to seek evidence of other sorts.

5. The evolution of mindreading and metacognition

The differing accounts outlined in section 1 lead to different commitments concerning the likely course of human evolution, and these in turn lead to different predictions about what we should expect to find in contemporary human cognition, and also in other species of animal. The present section shows that the "mindreading is prior" account comes out significantly ahead of its rivals in the former respect, before arguing that the animal data lend no support to either side.

All four of the accounts of the relationship between mindreading and metacognition can, and probably should, converge on essentially the same explanation of the evolutionary origins of human mindreading capacities. (Even those who think that mindreading capacities emerge in the course of childhood development through processes of learning that are akin to scientific theorizing insist that such theorizing has to begin with a specific innate basis; see Gopnik & Meltzoff 1997.) This will be some or other variant of the "Machiavellian intelligence" hypothesis (Byrne & Whiten 1988; 1997; Dunbar 2000), which points to the immense fitness advantages that can accrue to effective mindreaders among highly social creatures such as ourselves. And all should predict that one might expect to find simpler versions of mindreading capacity among other animals (perhaps confined to recognition of perceptual access and ignorance together with intention), especially among mammals who live in complex social groups. These predictions appear to be borne out (Call & Tomasello 2008; Cheney & Seyfarth 2007; Hare 2007; Hare et al. 2000; 2001; Tomasello et al. 2003a; 2003b).

Where the various accounts diverge is over the evolution of metacognition. From the perspective of a "mindreading is prior" account, no separate story needs to be told. Since metacognition, on this view, results from turning one's mindreading capacities upon oneself, its emergence will be a by-product of the evolution of mindreading. (This isn't to say that metacognition might not have come under secondary selection thereafter, perhaps by virtue of helping to build and maintain a positive self-image, as Wilson [2002] suggests.) All three competitor accounts, in contrast, have some explaining to do. This is most obvious in connection with a "two independent mechanisms" account of the sort championed by Nichols and Stich (2003). For if mindreading and metacognition are subserved by two (or more) cognitive mechanisms, then plainly there should be a distinct evolutionary story to be told about the emergence of each. But the same also holds in respect of a "one mechanism, two modes of access" account. Because neural connections are costly to build and maintain (Aiello & Wheeler 1995), some distinct evolutionary pressure will be needed to explain why the metarepresentational faculty (which might well have evolved initially for purposes of mindreading) should have acquired the input channels necessary to monitor the subject's own propositional attitudes.

The most natural way of explaining the structures postulated by the "metacognition is prior" account (championed by Goldman 2006) would likewise involve a distinct evolutionary pressure of some sort for the emergence of metacognition. The latter would happen first, followed subsequently by the integration of introspection with processes of imagination and simulative reasoning, presumably driven by the pressure to develop forms of "Machiavellian intelligence." Would it be possible to argue, however, that metacognitive capacities evolved to subserve mindreading from the start? It might be suggested that each incremental increase in metacognitive capacity was selected for because of its role in mindreading. For this account to work, however, it would have to be supposed that capacities to identify with others in imagination, together with dispositions to think and reason in simulation of the other within the scope of such a pretence, were already in place in advance of the appearance of both metacognition and mindreading. And one then wonders what such capacities would have been for. In the absence of any plausible suggestions, therefore, I shall assume that the "metacognition is prior" account, like the other two introspection-involving views, needs to postulate some evolutionary pressure in addition to those that issued in mindreading.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Thus, all three of the competitor accounts need to tell some story about the evolution of introspection. What I argue in section 5.1 is that the most popular such story – that metacognition evolved for purposes of self-monitoring and executive control of our own cognitive processes – makes predictions that are not borne out by the data. To the extent that this is true, then each one of those accounts is simultaneously disconfirmed. And this will therefore provide us with a further reason to accept the "mindreading is prior" account (in addition to the fact that it is the simplest, and should in consequence be accepted by default).

Although all three competitor accounts are committed to the existence of a distinct evolutionary pressure to explain the emergence of metacognition, only the "metacognition is prior" model makes a specific prediction about the *order* of emergence of the two capacities in phylogeny. It predicts, in particular, that we should be able to find metacognitive capacities in creatures that lack any capacity for mindreading (presumably because they lack the requisite imaginative abilities). Just this idea appears to motivate the recent flurry of interest in the metacognitive capacities of nonhuman animals (Terrace & Metcalfe 2005). This topic is examined in section 5.2.

5.1. The evolution of metacognition

What evolutionary pressures might have shaped the emergence of a distinct metacognitive capacity? One natural and very popular suggestion is that it was designed to have a supervisory role with respect to regular, first-order, cognitive processes – troubleshooting and intervening in those processes in cases of difficulty, initiating new strategies, checking that tasks are proceeding as expected, and so on (Shallice 1988). What I argue, however, is that although there is indeed a supervisory role for metacognition, it is one that does not require an introspective capacity distinct from the third-person mindreading system. I also argue that our metacognitive interventions are not capable of the sort of direct impact on cognitive processing that would be predicted if metacognition had, indeed, evolved for the purpose. But we first need to notice an important distinction.

Unfortunately, cognitive scientists use the term "metacognition" in two quite distinct ways, often without noticing the difference. (See Anderson & Perlis [2005a] for what seems to be a clear example. For distinctions related to the one drawn here, see Dennett 2000.) Generally the term is used, as it has been throughout this target article, to mean cognition *about* one's own cognition. Metacognition, in this sense, is inherently higher-order, involving metarepresentations of one's own first-order cognitive processes as such. But the word "meta" literally just means "above." And consequently many people understand metacognition to be any process that goes on *above* regular cognitive processes, performing a number of kinds of executive-function roles, such as monitoring the progress of a task and initiating new strategies when progress is blocked. On this view, any cognitive architecture that is organized into layers – containing not only a set of automatic information-generating and decision-making systems, but also a supervisory layer of some sort that can intervene in or alter the processes taking place in the first layer – will count as "metacognitive." But it is important to see that these supervisory processes needn't involve anything metacognitive in our first sense. For example, monitoring the progress of a task may just require a (first-order) representation of the goal-state, together with some way of comparing the current output of the system with the represented goal-state and making adjustments accordingly.

Indeed, all of the supervisory processes that Anderson and Perlis (2005a) describe as requiring both "self-awareness" and a "metacognitive loop" are actually just first-order processes organized into layers in this sort of way. For example, they describe a robot that is capable of noticing that it is no longer making forward progress (because it keeps bumping into a fence that it cannot see), and initiating an alternative strategy (e.g., traveling in an alternative direction for a while). There is plainly nothing metacognitive (in the sense of "metarepresentational") required here. The robot just needs to be on the lookout for failures to move forwards, and it needs to have been programmed with some alternative strategies to try when it doesn't. Even a mechanism that is capable of recognizing and responding to contradictions need only be sensitive to the *formal* properties of the representations involved, without representing them *as* representations. Thus, if representations of the form "P" and "∼P" are detected within active memory, the system might be programmed to place no further reliance on either of these premises, just as Anderson and Perlis suggest.

A significant portion of what gets described within cognitive science as "metacognition," then, should be set aside as irrelevant to the issues we are discussing. But of course a very large body of genuinely metacognitive data remains, especially in the domain of metamemory (e.g., Metcalfe & Shimamura 1994; Nelson 1992). But even where cognitive processes are genuinely metacognitive in the sense of being metarepresentational, deploying concepts of mental state types, they often operate without the capacity to intervene directly in the states and processes represented. For example, most metamemory capacities only require an ability to initiate or to intervene in *behavior*. Thus, a child might select one memorization task rather than another on the grounds that it contains fewer items (thus implicating knowledge *about* memory, but not intervening in the process of memory itself). Likewise, someone might mentally rehearse items in inner speech as an aid to memorization, which is an indirect behavioral influence on memory, not a direct intervention. And in the same spirit, it should be noted that while the intention to learn has an effect on study patterns, it has no effect on learning and recall once study patterns are controlled for (Anderson 1995). This is not what one would predict if metamemory were some sort of introspective capacity that had evolved for purposes of executive control, enabling subjects to intervene directly in the processes of memorization or memory retrieval. (Guiding behaviors that tend to issue in memorization or retrieval, in contrast, can be done equally well by a mindreading system.)

Koriat et al. (2006) review much of the extensive literature on metamemory, and experimentally contrast two competing models. One is that metacognitive monitoring serves the function of controlling and directing the underlying cognitive processes. (Plainly this would be consistent with the evolutionary explanation of introspection

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

sketched earlier.) The other is that metacognitive judgments are evidence-based, cued by experiences that are caused by the cognitive processes in question. (This would be consistent with the self-interpretative position being developed here.) Although they do find metacognitive phenomena that fit the former profile, none of these suggests any real role for introspection of attitudes. Rather, they include such phenomena as allocating greater study time to items that attract a larger reward. In contrast, there is extensive evidence of cue-based metacognitive judgments. Thus, feelings of knowing are often based on the ease with which one can access fragments of the target knowledge (Koriat 1993) or items related to the target (Schwartz & Smith 1997). And judgments of learning made during or after study are based on the "fluency" with which items are processed during study itself (Begg et al. 1989; Benjamin & Bjork 1996; Koriat 1997). Again, this isn't at all what one would predict if one thought that a capacity for introspection of attitudes had evolved for purposes of metacognitive control. For why, in that case, would one *need* to rely on indirect cues of learning?

While the influence of metacognitive judgments on cognitive processes is often indirect, it should be stressed that such judgments are actually intrinsic to the sorts of processes that would be characterized as belonging to "System 2," as we will see in section 7. Human beings sometimes engage in forms of conscious thinking and reasoning that are thoroughly imbued with metacognitive beliefs and judgments. But what appears to make such forms of thinking consciously accessible is that they are conducted in inner speech and other kinds of imagery. In which case the type of metacognitive access that we have, here, will turn out to be fully consistent with a "mindreading is prior" account.

The preliminary upshot of this discussion, then, is that the predictions generated by the most common evolutionary explanation of an introspective capacity (namely, that its purpose is executive monitoring and control) are not borne out by the data. This provides us with good reason to embrace the alternative "mindreading is prior" account instead.

5.2. Metacognitive processes in nonhuman animals

The last few years have seen a flurry of experiments purporting to demonstrate the presence of metacognitive processes in nonhuman animals (Beran et al. 2006; Call & Carpenter 2001; Hampton 2001; 2005; Hampton et al. 2004; Kornell et al. 2007; Shields et al. 1997; Smith 2005; Smith et al. 1995; 1997; 2003; Son & Kornell 2005; Washburn et al. 2006). If these experiments were to prove successful, and if the animals in question were to lack any capacity for mindreading of attitudes (as most researchers assume), then this would provide dramatic support for the view that metacognition is prior to and underpins mindreading. (By the same token, it would provide powerful evidence *against* the "mindreading is prior" account being defended here.) These studies are reviewed and critiqued in detail in Carruthers (2008b), where I demonstrate that all of the phenomena in question are readily explicable in first-order terms. Here I shall confine myself to outlining my treatment of just one of the simpler alleged instances of animal metacognition.

Smith et al. (2003) argue that the adaptive behavioral choices made by monkeys and dolphins in conditions of uncertainty demonstrate that the animals are aware of their own state of uncertainty and are choosing accordingly. Thus, monkeys who have been trained to discriminate between dense and sparse visual patterns, and to respond differentially as a result, will increasingly make use of a third "don't know" option (which advances them to a new trial without the penalty of a delay) when the patterns are made harder and harder to distinguish. But all that is really needed to explain the animals' behavior here is an appeal to *degrees* of belief and desire. For an animal that has a weak degree of belief that the pattern is dense and an equally weak degree of belief that the pattern is sparse, will have correspondingly weak and balancing desires to make the "dense" response as well as to make the "sparse" response. In contrast, the animal will have a high degree of belief that the "don't know" response will advance to a new trial without a timeout, and a timeout is something that the animal wants to avoid. Hence, pressing the "don't know" key will be the strongest-motivated action in the circumstances. No metacognitive forms of awareness of the animal's own mental states are required.

Of course humans, when they have performed tasks of this sort, will report that they were aware of a feeling of uncertainty, and will say that they chose as they did *because* they were uncertain. There is no problem here. Although these reports are metacognitive, and reflect metacognitive awareness, the processes reported on can be first-order ones, just as they are for the monkeys. In both species uncertainty will be accompanied by feelings of anxiety, which will motivate various forms of information-seeking behavior (such as moving one's head from side to side for a better view), as well as a search for alternatives. But humans, with their highly developed mindreading capacities, will interpret these feelings and resulting behaviors for what they are – manifestations of uncertainty. It is only if a human reports that she acted as she did, not just because she *was* uncertain, but because she was *aware of being* uncertain, that there will be any conflict. Such reports are likely to be false, in my view. For the most part the "executive function" behaviors that we share with other animals are best explained in terms of the first-order processes that we also share (Carruthers 2008b). It is only when we consider forms of behavior that are unique to humans that we need to appeal to metacognitive processes.[7] But these can all be processes that I shall describe in section 7 as belonging to "System 2", which don't require any faculty of introspection distinct from mindreading.

6. The confabulation data

There is extensive and long-standing evidence from cognitive and social psychology that people will (falsely) confabulate attributions of judgments and decisions to themselves in a wide range of circumstances, while being under the impression that they are introspecting (Bem 1967; 1972; Eagly & Chaiken 1993; Festinger 1957; Nisbett & Wilson 1977; Wegner 2002; Wicklund & Brehm 1976; Wilson 2002). These data are consistent with a "dual method" account of metacognition (Goldman 2006), according to which metacognition is

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

sometimes self-interpretative and sometimes introspective. But given that we have been offered, as yet, no positive reasons to believe in the reality of introspection for attitudes, the best explanation at this stage is that metacognition *always* results from people turning their mindreading abilities upon themselves.

Literally hundreds of different studies have been conducted charting confabulation effects and the circumstances under which they occur; and a number of different explanatory frameworks have been proposed ("cognitive dissonance," "self-perception," and others). I have space only to describe a few salient examples and to discuss some of the ways in which an introspection theorist might attempt to respond.

First, however, let me mention some types of confabulation data that *aren't* relevant for our purposes. One emerges from studies that find people to be inaccurate in reporting the *causes* of their judgments or behavior. For example, people are notoriously bad at identifying the factors that persuade them of the truth of a message or the quality of a job interviewee. Such cases raise no difficulty for a believer in introspection. The reason is simple: no one thinks that causation can be introspected. It is supposed to be the *occurrence* of our attitudes that is accessible to introspection, not the causal role (if any) that those attitudes have in any given situation. This could only be known by theorizing. Likewise, we should set to one side studies in which subjects are required to report on their attitudes some significant time afterwards. Thus, the fact that subjects will, at the end of the experiment, confabulate lesser enjoyment in playing with a game when they had been paid to play with it (belied by the amount of time that they had freely devoted to the game in their spare time; Kruglanski et al. 1972) raises no difficulty for an introspection theorist. For, given the proposed on-line monitoring function for introspection, it makes sense that no medium- or long-term record of introspected mental events should normally be kept. And in the absence of any such record, subjects have no option but to self-interpret. (The cognitive monitoring account must require that brief records of introspected events should be kept in some sort of working memory system, however. So we should expect subjects to be capable of giving introspective reports for a few moments after the events have occurred. This point is relevant to a number of the experiments described below.)

Now consider one of the classic studies conducted by Nisbett and Wilson (1977). Subjects chose between four items of panty-hose (which were actually identical), thinking that they were taking part in a market survey. They displayed a strong right-hand bias in their choices, but all offered judgments of quality ("I thought that pair was the softest," etc.) immediately afterwards in explanation of their choice. Nisbett and Wilson themselves cast this result in terms of confabulation about the *causes* of action, and those who believe in the introspectability of judgments will often dismiss it on that ground (Rey 2008). But this is to miss the point that subjects are *also* confabulating and attributing to themselves a *judgment* (albeit one they believe to have caused their action, and at least on the assumption that they didn't *actually* judge the right-hand item to be softest – otherwise the first-order mechanisms discussed in sect. 2.1 could underlie their reports). How could one claim otherwise? Well, it is likely that the root cause of the right-hand choice bias is a right-hand *attention* bias, and someone might claim that attending more to the right-hand items causes subjects to judge that those items are softer (or are of better quality, or a nicer color, etc.). These judgments can then be introspected and veridically reported. But the causal pathways postulated here are pretty mysterious. And the most likely candidates for fleshing them out are ones that already involve confabulation. (For example, noticing that I am attending more to the right-hand item, and noticing that it is soft, my mindreading faculty might hypothesize that I am paying more attention to it *because* it is the *softest*, leading me to ascribe to myself just such a judgment.)

There is also ample evidence of confabulation for decisions. For example, Brasil-Neto et al. (1992) caused subjects to move one index finger or another via focal magnetic stimulation of areas of motor cortex in the relevant brain hemisphere. (Subjects had been instructed to freely decide which finger to move when they heard a click, which was actually the sound of the magnet being turned on.) Yet the subjects themselves reported *deciding* to move that finger. Now, it is very unlikely that stimulation of motor cortex should itself cause a decision (as well as causing movement), hence giving rise to a propositional attitude event that can be introspected. For if the back-projecting pathways between motor cortex and frontal cortex were used for this purpose, then one would predict that stimulation of premotor cortex would also have such an effect; but it does not (Brasil-Neto et al. 1992).

Further evidence of confabulation for decisions is provided by Wegner and Wheatley (1999), who induced in subjects the belief that they had just previously taken a decision to stop a moving cursor on a screen (which was controlled via a computer mouse operated jointly with a confederate of the experimenter), by the simple expedient of evoking a semantically relevant idea in the subject just prior to the time when the confederate actually caused the cursor to stop. (Subjects heard a word through headphones, ostensibly as a distracter, shortly before the confederate was able to bring the cursor to a stop beside a picture of the named object.) It seems that the subject's mindreading faculty, presented with the evidence that the subject had been thinking of the relevant object shortly before the cursor came to a stop beside it, reasoned to the most likely explanation, and concluded that the subject had taken a decision to stop beside that very object. (A control condition ruled out the possibility that hearing the semantically relevant word caused an actual decision to stop the cursor next to the named object.)

It might be objected that all of the examples considered so far are ones where (plausibly) actually no judgment was made, nor any decision taken, although behavior occurred that led subjects to think that it had. Hence, someone might propose that it is only in such cases that confabulation occurs. Whenever there *is* a propositional attitude event, it might be said, it can be introspected; and only when there isn't, will subjects self-interpret. However, if there really were two distinct ways of attributing judgments and decisions to oneself (an introspective mode as well as an interpretative one), then it would be odd that the latter should always win out in cases where no judgment or decision has actually been made. For presumably

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

an introspective mechanism can detect an absence. And if the introspective mechanism is delivering the judgment, "No judgment," or, "No decision" at the same time as the mindreading system is attributing one to oneself, then why is it that the latter should always dominate, leading to confabulated answers to the experimenters' questions? On the contrary, since the introspective mechanism is supposed to have evolved to be especially direct and reliable, one would expect it to be routinely given precedence in cases of conflict.

Consider some further data: Subjects who emerge from a hypnotic trance, and then later carry out an instruction given to them while hypnotized, will often confabulate an explanation for their action (Sheehan & Orne 1968; Wegner 2002). Presumably what happens is that they decide, while hypnotized, to comply with the request of the hypnotist. And the effect of this decision is to set up a conditional intention – for example, "When I see the book on the table I shall place it on the shelf" – which remains in existence once the hypnotic episode and original decision are forgotten. This intention is then activated thereafter when the antecedent of the intention is fulfilled (e.g., the book is seen). In which case, there *is* a decision here to report. And if the subject were to confine herself to reporting just that decision (e.g., to put the book on the shelf), then she would report veridically. But in fact she confabulates a further judgment and/or goal – for example, that the book is out of place and makes the room look untidy.

It might be said in reply that placing a book on a shelf isn't something that people normally do for its own sake. Hence, there are powerful pragmatic reasons for the agent to confabulate a further attitude when pressed by the experimenter to explain her action, even given that the introspective mechanism is detecting the absence of any such state (Rey 2008). But this explanation is problematic. For there are all sorts of circumstances in which people are perfectly content to say, "I don't know why; I just did it" when asked to explain why they acted in a particular way. Why should the same not be true here? Indeed, it isn't uncommon to catch oneself performing actions of precisely this sort – absent-mindedly moving a household item from one place to another – in circumstances where one is prompted to ask oneself, "Why did I just do that?", or where one replies if challenged for an explanation, "I don't know; just a nervous tic I suppose." But in any case, Rey's suggestion should be testable: The hypnotist could instruct a subject to perform a movement that is ambiguous between two distinct actions (e.g., greeting someone with a wave versus waving away a bug), one of which is very much more likely in the circumstances (e.g., indoors, occurring just as someone known to the subject enters the room). The hypnotist's instruction would be formulated in terms of the less likely action: "When John enters the room you will raise your arm and move it back and forth with the palm facing forwards to shoo away any bugs." On Rey's introspective account, subjects should offer the latter in explanation of their arm movement. A "mindreading is prior" theorist will predict, in contrast, that subjects should offer the more likely explanation: "I was waving to John."

There is also an extensive and long-standing set of data that subjects' behavior, when caused in ways that they are unaware of or inattentive to, will lead them to confabulate when describing their own degree of belief in some proposition. (See Bem 1967; 1972; Cooper & Duncan 1971; Festinger 1957; Greenbaum & Zemach 1972; Wicklund & Brehm 1976; see also, Eagly & Chaiken 1993 for a more recent review.) Thus, subjects who are manipulated into writing a counter-attitudinal essay for meager pay, but who do so believing that they have made a free decision, will say that they have a greater degree of belief in the proposition that their essay was defending than will subjects in the same circumstances who are paid a decent sum of money. It seems that subjects reason: "I wrote the essay freely, but I can't have done it for the money, so I must believe it." And indeed, subjects who don't participate, but have the circumstances of the various essay-writers described to them, make just such an inference.

Likewise, it has long been known that subjects who are induced to nod their heads while listening to a tape via headphones (ostensibly to test the headphones themselves) will say that they have a greater degree of belief in the propositions being defended on the tape than will subjects who are induced to shake their heads (Wells & Petty 1980). It seems that subjects reason: "Since I am nodding/shaking my head, this is evidence that I believe/disbelieve the propositions asserted." Admittedly, this is not the only explanation possible. It might be that head-nodding primes for positive thoughts about the message, which in turn cause greater agreement, which is then veridically reported. Briñol and Petty (2003) set out to test this alternative by varying the persuasiveness of the messages themselves. When the message is persuasive, nodding increases belief and head-shaking decreases it, which is consistent with either one of the two explanations. But when the message is *un*persuasive, the opposite occurs: nodding *decreases* belief and head-shaking *increases* it. The authors present evidence that what is actually happening is that subjects interpret their own nodding behavior as confirming their own initial negative reactions to the message, whereas head-shaking is interpreted as disagreement with those reactions.

Now, it does not *follow*, logically, from all this (and much more) data that there is no such thing as introspection for propositional attitudes. For there might be one set of such events to which we have introspective access while there is another set that we can't introspect; and hence, whenever our behavior is caused by attitudes drawn from the latter set, we are forced to self-interpret (and often to confabulate). What might be proposed, in effect, is that there is both a conscious and an unconscious mind. Judgments and decisions within the conscious mind are introspectable, whereas judgments and decisions within the unconscious mind can only be known (if at all) by turning our mindreading capacities upon ourselves. And just such a view seems to be endorsed by some of those who have been most prolific in demonstrating the reality of metacognitive attitude attribution via processes of interpretation and confabulation. Thus, both Wegner (2002) and Wilson (2002) allow that we do sometimes have introspective access to our (conscious) thoughts, even if much of the time our access to our own propositional attitudes is interpretative, and often confabulatory.

In order for this proposal to count as a realistic competitor to the interpretation-only alternative, however, we need some principled account of the two forms of mentality and their relationships to each other. This isn't by any

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

means an easy thing to provide. For we need to know what it is about some judgments and decisions that makes them available for introspection, while others are cut off from such availability. What kind of cognitive architecture can underlie and explain these patterns of availability and unavailability in anything more than an ad hoc way? I take up this challenge in the next section, where the only such account that I know of is outlined and discussed. It will turn out on closer investigation, however, that the account actually lends no support to the introspectionist position.

7. Is there a conscious mind?

One possible response to our challenge is to distinguish between two different *levels* of mental process (conscious and unconscious). And the only worked-out account of these two levels that I know of is as follows. It would be allowed that the access we have to our unconscious attitudes (whether or not they get expressed in speech or other imagery) is always interpretative, as argued earlier. But it might be claimed that the stream of inner speech and other forms of imagery is constitutive of a distinct kind of (conscious) mentality (Frankish 2004). Certainly such events are not epiphenomenal, but they often make an important causal contribution to subsequent thought and behavior (Carruthers 2002; 2006; Clark 1998). And it might be said that such events are routinely available to introspection.

This suggestion comports very naturally with an idea that has been gaining increasing ground among those who work on the psychology of reasoning (Evans & Over 1996; Kahneman 2002; Sloman 1996; 2002; Stanovich 1999). This is that human reasoning processes may be divided into two very different types, often now referred to as "System 1" and "System 2. System 1 (which is really a set of systems, arranged in parallel) is fast, unconscious, hard to alter, universal to all thinkers, and evolutionarily ancient. System 2, in contrast, is slow and serial, characteristically conscious, malleable in its principles of operation, admits of significant variations between individuals, and is evolutionarily novel. And a number of authors have emphasized the important constitutive role played by imagery (especially inner speech) in the operations of System 2 (Carruthers 2009; Evans & Over 1996; Frankish 2004). Likewise, others have demonstrated the crucial role played by inner speech in the performance of tests of executive functioning (which are likely to implicate System 2), such as the Wisconsin Card Sorting Task (Baddeley et al. 2001). For when inner speech is suppressed by the need to shadow an irrelevant speech stream while performing the task, performance collapses.

In order for this account to be successful, however, it is obviously crucial that the conscious imagistic events in question should play the right sorts of causal role, constitutive of the roles of the various attitude types. Not any old causal role will do. Thus, it is a conceptual constraint on an event being an instance of *deciding*, for example, that it should fit one of two causal profiles (Bratman 1987; 1999). In the case of a decision to act here-and-now, the decision should issue in motor instructions without the intervention of any further practical reasoning. A decision is supposed to *end* the process of practical reasoning and to *settle* what I do (unless something goes awry with my motor system, of course). Something similar is true of a decision to act in the future: this should settle *that* I act (unless something significant changes in the interim) and *what* act I shall perform. Any further reasoning in the future should be confined to the question of *how* to act. Intentions for the future place constraints on our practical reasoning. They have the form of partial plans, in which details may be left blank to be filled in later, but in which the overall structure is fixed.

A similar point can be made about judgments. Just as a decision is an event that ends a process of practical (action-oriented) reasoning, so a (non-perceptual) judgment is an event that concludes a piece of theoretical (belief-oriented) reasoning. A judgment, then, is an event that will normally (a) immediately (without further inference) give rise to a stored standing-state belief with the same content, and (b) will immediately be available to inform practical reasoning, interacting with the subject's goals (where appropriate) in the construction of plans. If an event is genuinely a judgment, then there should be no further cognitive activity standing between it and the normal roles of judgment (the formation of belief and the guidance of action).

We need to ask, therefore, in what way is it that the events that constitute System 2 achieve their characteristic effects. For only if these events have the right sorts of causal roles, can they be said to *be* propositional attitude events of judging, deciding, and the like. And so, only if they have the right sorts of roles, can our introspective, non-interpretative, awareness of them (which I grant) constitute introspective, non-interpretative, awareness of a set of propositional attitudes.

The processes that take place in System 2 don't simply mirror those that take place in System 1, of course, tracking them one-for-one. Rather, sequences of imagery can occur in accordance with well-practiced rules or habits, or they can be guided by subjects' beliefs about how they *should* reason, often issuing in an assertoric statement, for example, that isn't simply the expression of a pre-existing (System 1) judgment.[8] So let us consider such a case. As a result of an episode of System 2 conscious activity, I might formulate and rehearse the assertoric utterance, "Polar bears are endangered." Under interpretation, this event will likely be heard as an assertion that polar bears are endangered. And as a result, I will think and act in the future much as if I had formed just such a judgment. I shall, for example, reply positively if asked whether or not polar bears are endangered. And if one of my goals is to try to protect endangered species, then I might, in consequence of this event, begin writing a suitable letter to my congressional representative.

How does the rehearsed assertion achieve these effects? There are a number of possibilities. (These aren't mutually exclusive, I should stress. On the contrary, a pluralist position concerning the realization of System 2 processes is probably correct; see Carruthers 2009.) One is that the event causes me to believe of myself (unconsciously, at the System 1 level) that I believe polar bears to be endangered. Then this, together with a standing desire to think and act consistently, will lead me to answer positively when asked whether or not I believe that polar bears are endangered. And it might also issue in letter-writing behavior. For if I believe myself to believe that polar bears are

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

endangered, and want to do something to help endangered species, then consistency requires that I should act.

Another possibility is that my mentally rehearsed assertion causes me to believe I have committed myself to the truth of the proposition that polar bears are endangered. And then a standing (System 1) desire to execute my commitments will lead me to act in ways I consider to be appropriate to that commitment. And yet another possibility is that the rehearsed sentence is treated by my cognitive systems much as if it were an item of testimony from a putatively reliable informant, and after checking for coherence with existing belief, it is then stored as a first-order (System 1) belief, which then issues in appropriate behavior in the normal way.

The important point to notice is that on each of these three accounts, the rehearsal of the assertion "Polar bears are endangered" does *not* give rise to a standing-state belief immediately, without the mediation of any further cognitive processing. Nor is it immediately available to guide planning with respect to endangered species. For in each case further, down-stream, cognitive activity must occur first. Either I must form the belief that I believe polar bears to be endangered, which then interacts with a higher-order desire to guide activity consistent with my possessing such a belief. Or I must form the belief that I have made an appropriate commitment, which again has to interact with a higher-order desire to execute my commitments in order to guide behavior. Or the assertion must be evaluated in something like the way that the testimony of other people is (checking for coherence with existing belief, and so on; see Harris [2002a; 2002b] who shows that even young children don't automatically accept the testimony of others, but evaluate it in light of a variety of "gate-keeping" criteria first). In each of these cases the relevant assertion does *not* have the right sort of causal role to be a judgment. For it does not by itself settle what I believe.

An exactly parallel argument can be constructed for System 2 episodes that might be candidate decisions, such as saying to myself (in inner speech) at the conclusion of a period of System 2 activity, "So, I shall write to my congressional representative." This utterance does not, by itself, settle anything. For it first has to give rise to the belief that I have decided to write, or to the belief that I have committed myself to write, and then the causal pathways operate as described. So in each case, then, although there is a conscious System 2 event to which I have introspective access, it *is not* an event of deciding on an action, or of forming a new judgment. And this argument generalizes to other candidate types of propositional attitude, such as *supposing* something to be the case, or *fearing* that something is the case, and so forth.

(Interestingly, however, System 2 conscious activity *is* constitutive of *thinking*. For there are few significant conceptual constraints on what sorts of processes can count as thinking. Roughly speaking, any sequence of content-bearing events that makes some difference to subsequent attitude-formation or to behavior can count as thinking. So we *do* have introspective access to some forms of thinking – specifically to imagistically expressed System 2 thinking – even if, as I have argued, we don't have such access to any propositional *attitudes*.)

I conclude there is, indeed, such a thing as conscious mentality. In addition to globally broadcast experiences of various sorts, there are also sequences of visual and auditory imagery that make an important difference to our cognitive and practical lives. But our introspective access to these events doesn't thereby give us introspective access to any propositional attitudes. On the contrary, our only form of access to propositional attitudes of judging, deciding, and so forth, is interpretative.

8. The evidence of unsymbolized thinking

Recall from section 1 that a "mindreading is prior" account makes two distinctive predictions. The first is that it should be possible for subjects to be misled, in attributing propositional attitudes to themselves, by being presented with manipulated behavioral or sensory data. As we have seen in sections 6 and 7, this prediction is amply confirmed, in ways that the opposed accounts cannot easily accommodate. But the second prediction is that subjects should be incapable of attributing propositional attitudes to themselves in the *absence* of behavioral or sensory data. All three of the opposing positions, in contrast, make the opposite prediction. Because they maintain that introspection for propositional attitudes exists, subjects should generally have no need of evidence of any kind when making self-attributions. The presence of behavioral and sensory cues should be entirely accidental. However, we have already seen in section 5.1 that many kinds of metacognitive judgment – such as judgments of learning – are actually dependent upon sensory cues. Hence, in these cases, at least, the sensory cues *are not* accidental. The present section evaluates some additional evidence that bears on this matter.

The data in question derive from "introspection sampling" studies conducted with normal subjects, using the methodology devised by Hurlburt (1990; 1993). Subjects wear a paging device throughout the day, via which they hear a "beep" at randomly generated intervals. Subjects are instructed to "freeze" the contents of their consciousness at the very moment of the beep, and to make notes of it, to be discussed and elaborated in a later meeting with the experimenter. All normal subjects report, in varying proportions, the occurrence of inner speech, visual imagery, and emotional feelings. But many subjects also report the presence of "purely propositional," unsymbolized thoughts at the moment of the beep. In these cases subjects report thinking something highly determinate – such as wondering whether or not to buy a given box of breakfast cereal – in the absence of any visual imagery, inner speech, or other sensory accompaniments.

So far there isn't any difficulty here for a "mindreading is prior" account. For such an account doesn't have to claim that all thinking should be imagistically expressed. Indeed, quite the contrary: the thoughts generated by the mindreading system itself will characteristically remain *un*expressed. What the account does claim is that self-attributions of thought should be dependent on the presence of either sensory/imagistic *or behavioral/circumstantial* data. And what is striking about a good many of the instances of self-attributed unsymbolized thought is that they occur in circumstances in which a third-party observer might have made precisely the same attribution. If you saw someone standing motionless, looking reflectively at a box of breakfast cereal on a supermarket shelf, for

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

example, you might well predict that she is wondering whether or not to buy it. Our suggestion can therefore be that when prompted by the beep, subjects turn their mindreading systems on their own behavior and circumstances (together with any sensory or imagistic cues that might be present), often enough interpreting themselves as entertaining a specific thought. Provided that the process happens swiftly, this will then be self-attributed with all of the phenomenological immediacy and introspective obviousness as normal.

Although a great many of the examples in the literature can be handled in this way, not quite all of them can. For instance, at the time of the beep, one subject reported that she was wondering whether her friend who would be picking her up later that day would be driving his car or his truck. This thought seemed to occur in the absence of any inner speech or visual imagery. Yet there was nothing in the subject's immediate circumstances or behavior from which it could be derived, either. What cannot be ruled out, however, is that the thought in question was self-attributed because it made the best sense of sensory activity that had been taking place just *prior* to the beep – for example, two memory images deriving from previous experience, in one of which the friend arrives in his car and in the other of which he arrives in his pickup truck. Since Hurlburt's methodology makes no provision for collecting data on experiences occurring shortly before the beep, we simply don't know. An extension of the methodology might provide us with a valuable test, however. Another possible test would be to look for correlations between the extent to which different subjects report purely propositional thoughts (with quantities of inner speech and visual imagery controlled for) and the speed of their mindreading abilities in third-person tasks. Because subjects will only have the illusion of introspecting if they can reach a self-interpretation smoothly and swiftly, I predict that there should be a positive correlation.

Hurlburt and Akhter (2008) concede that it is possible that attributions of unsymbolized thought to oneself might result from swift and unconscious self-interpretation. But they present the following consideration against such a possibility. Subjects are initially quite reluctant and hesitant in describing instances of unsymbolized thought, presumably because they share the commonly held folk theory that all conscious thinking is accompanied by images of one sort or another. But explicitly held folk theories are one thing; assumptions built into the operations of the mindreading faculty are quite another. And there is no reason to think that the latter will share all of the culturally developed assumptions made by the folk. Hence, the mindreading system might have no hesitation in attributing a thought to the self in the absence of any sensory cues, even though the person in whom that system resides does so hesitate. I conclude this section, therefore, with the claim that although there is no *support* to be derived for a "mindreading is prior" account from the introspection-sampling data, neither is there, as yet, any evidence to count against it.

9. The evidence from schizophrenia

Recall from section 1 that two of the three competitor models (namely, models 1 and 2) predict that there should exist cases in which mindreading is intact while metacognition is damaged. The "mindreading is prior" account, in contrast, must deny this. Nichols and Stich (2003) cite certain forms of schizophrenia as confirming the former prediction. More specifically, patients with "passivity" symptoms, who claim that their own actions are not under their control and that their own episodes of inner speech are somehow inserted into their minds by other people, are supposed to demonstrate such a dissociation (presumably on the grounds that such patients no longer have normal introspective access to their own behavioral intentions).[9,10] This is because such patients perform normally when tested on batteries of mindreading tasks.

There is no reason to think that the symptoms of passivity forms of schizophrenia are best explained by a failure of metacognitive competence, however. Rather, the damage lies elsewhere, resulting in faulty data being presented to the mindreading system. Frith et al. (2000a; 2000b) provide a detailed account designed to explain a range of disorders of action and awareness of action (including passivity-symptom schizophrenia). The account builds on well-established models of normal action control, according to which an "efference copy" of each set of motor instructions is transformed via one or more body emulator systems and used to construct a "forward model" of the expected sensory consequences of the movement. This can then be compared, both with the motor intention itself and with the incoming perceptual data, allowing for swift correction of the action as it unfolds (Grush 2004; Wolpert & Ghahramani 2000; Wolpert & Kawato 1998). Frith et al. think that the symptoms of passivity and "alien control" in schizophrenia can be explained as issuing from damage to this action-monitoring system, which results in no forward model ever being created for comparison.

Now the important point to note for our purposes is that the kind of action-monitoring just described is entirely first-order in character, and qualifies as "metacognitive" only in the weak and irrelevant sense distinguished in section 5.1. There is no reason to think that it should involve metarepresentions of our own motor intentions, let alone introspective access to them. And indeed, the speed with which the monitoring process operates suggests very strongly that introspection *is not* involved (Jeannerod 2006).

But why should the absence of a forward model lead subjects to feel that their actions are not their own? Frith et al. (2000a) point out that the forward model is normally used to "damp down" experiences resulting from movement that are of the sort predicted in the forward model. This is why it is normally impossible to tickle yourself, whereas if you wear special gloves that introduce a slight delay in your movements, then self-tickling suddenly becomes possible (Blakemore et al. 1998; Weiskrantz et al. 1971). And it is also why when you unwrap a candy at the opera you barely hear it while those around you are disturbed. If no forward model is created, however, then perceptions resulting from your actions will be experienced with full vividness, just as if the movements had been caused by another person. The suggestion is that passivity-symptom schizophrenics have the sense that their actions are caused by others because those actions literally *feel* that way to them.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

In addition, one might expect the comparator process to give rise to heightened attention and feelings of anxiety in cases where there is too great a mismatch between the forward model and the perceptual data received. These feelings would be especially enhanced in cases where there is *no* forward model, as a result of some pathology. For the comparator system would be receiving perceptual input of an action being performed, but without receiving the normally attendant input deriving from an efference copy of a motor intention. So this would, as it were, be a case of maximum mismatch. An additional suggestion, then, is that these feelings of anxiety might signal to the mindreading system that something is amiss, perhaps reinforcing the impression that the actions are not one's own. Put differently: Only when everything is going smoothly, with no feelings of anxiety or surprise specifically attending one's action, does the mindreading system attribute agency to the self by default.

I conclude that passivity-symptom forms of schizophrenia are not best interpreted as instances of a dissociation between mindreading and metacognitive capacities. Rather than being cases in which mindreading is intact while introspection is damaged, the damage is to lower-level forward modeling and/or comparator systems. This results in experiences that are naturally interpreted as indicating that one's actions (including one's mental actions, such as inner speech) are not one's own.

10. The evidence from autism

The final major area in which the relationship between mindreading and metacognition can be assessed concerns autism. Almost everyone agrees that third-person mindreading is significantly impaired in autism. (There is, however, disagreement over whether this impairment lies at the heart of the syndrome.) In which case the prediction of a "mindreading is prior" account will be that autistic people's access to their own propositional attitude states must be impaired as well. Nichols and Stich (2003) and Goldman (2006) each maintain, in contrast, that introspection is intact in autism, with difficulties in other-understanding arising from difficulties in supposing or empathizing.

One set of data concerns an introspection sampling study conducted with three adult autistic men (Frith & Happé 1999; Hurlburt et al. 1994). All three were able to report on what was passing through their minds at the time of a randomly generated "beep," although one of them experienced significant difficulties with the task. This is interpreted as demonstrating that introspection is intact in autism. There are two points to make. First, none of these three subjects was entirely deficient at mindreading. On the contrary, two of them could pass second-level false-belief tasks, and the third could pass simple first-level false-belief tasks. So no one should predict that any of them would be entirely deficient at self-attribution, either. (It is worth noting, moreover, that the experimenters found a strong correlation between the subjects' abilities with third-person tasks and the sophistication and ease of their introspective reports. This finding is problematic for the view that introspection is undamaged in autism.) Second, the form of "mindreading is prior" account being defended here predicts that people with autism should have no difficulty in reporting the occurrence of perceptions, images, or emotional feelings, provided that they possess the requisite concepts. For these events will be globally broadcast and made directly accessible to their (damaged but partially functioning) mindreading faculties. And indeed, much of the content of the introspective reports of the three autistic subjects concerned visual imagery and emotional feelings. Reports of their own occurrent attitudes tended to be generic ("I was thinking . . ."), and one of the three men (the one who could only pass first-level false-belief tasks) had significant difficulties in reporting his own attitudes at all.

Another set of data of the same general sort concerns the autobiographical reports of adults with autism, who are often able to describe with some vividness what their mental lives were like at ages when they almost certainly wouldn't have been capable of attributing mental states to other people. Nichols and Stich (2003) comment that (provided we accept the memory reports as accurate), the individuals in question must have had reliable introspective access to their own mental states prior to having any capacity for mindreading. But actually we have no reason at all to believe that memory is itself a second-order (metarepresentational) process. When I observe an event, a first-order representation of that event may be stored in memory. When that memory is later activated, I shall describe it by saying that I remember *seeing* the event in question (say). But it doesn't at all follow that the original event involved any metarepresentation of myself as seeing something. Likewise for other sorts of memories and other sorts of mental events. The fact that autistic adults give metarepresentational reports of their mental lives as children does not show that autistic children are capable of metarepresenting their own mental states. It just shows that they are capable of memory formation.

Nichols and Stich (2003) also place considerable reliance on a study by Farrant et al. (1999), who tested autistic children, as well as learning-disabled and normal children matched for verbal mental age, on a range of metamemory tasks. Since they were able to find no significant differences between the groups, the authors conclude that metacognition is unimpaired in autism. Two preliminary points should be emphasized about this study, however. One is that almost all of the autistic children tested were sufficiently well advanced to be able to pass first-order false-belief tasks. So we should predict that they would have some understanding of their own minds, and that they would be capable of completing simple metacognitive tasks. Another point is methodological: The small group sizes meant that statistically significant differences were not detected even when a trend (namely weaker performance by the autistic children) was plainly visible in the raw data. We simply don't know whether those trends would have been significant had larger groups of children been used.

A deeper problem with the Farrant et al. data is that none of the experimental tasks was metacognitive in the right sort of way, requiring access to the subject's current propositional attitudes. On the contrary, they could be solved by anyone who possessed the requisite mental concepts who was also a smart behaviorist. For example, one experiment tested whether children with

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

autism were aware that it is easier to learn a small number of items than a larger number. Not surprisingly, the children did well on this test. But they would have had ample opportunity over a number of years of schooling to have established a reliable correlation between the number of items studied in a task and the number of responses later given that are evaluated as correct. (Note that the average age of the autistic children in this experiment was 11 years.)

It is true that many of the autistic children in question could give simple verbal descriptions of some memorization strategies. But many of these involved such tasks as looking in likely places (for an object that had been mislaid) or listening carefully to the instructions (from someone reciting a list of things to remember). This is metacognitive only in the minimal sense of mentioning looking and listening. Moreover, in order to develop a cognitive strategy like mental rehearsal (which a number of the autistic as well as normal subjects suggested), it is doubtful that much mindreading ability is required. Rather, children just need to notice a positive correlation between a behavior (rehearsal) and an outcome (getting the correct answer), which should be well within the reach of even a clever behaviorist (provided that the latter had access also to *inner* behavior, such as inner speech).

Thus, the data from autistic people considered by Nichols and Stich (2003) and by Goldman (2006) do not support their introspectionist positions against an interpretative, "mindreading is prior" account. But there are other data that these authors don't discuss, which suggest that people with autism are decidedly poor at attributing propositional attitudes to themselves. Let me describe just a couple of strands of evidence here.

Phillips et al. (1998) tested children with autism against learning-impaired controls (matched for verbal mental age) on an intention reporting task. The children had to shoot a "ray gun" at some canisters in the hopes of obtaining the prizes contained within some of them. But the actual outcome (i.e., which canister fell down) was surreptitiously manipulated by the experimenters (in a way that even adults playing the game couldn't detect). They were asked to select and announce which canister they were aiming at in advance (e.g., "The red one"), and the experimenter then placed a token of the same color next to the gun to help them remember. After learning whether they had obtained a prize, the children were asked, "Did you mean to hit that [for example] green one, or did you mean to hit the other [for example] red one?" The autistic children were much poorer than the controls at correctly identifying what they had intended to do in conditions where there was a discrepancy between intention and goal satisfaction. For example, if they didn't "hit" the one they aimed at, but still got a prize, they were much more likely to say that the canister that fell was the one they had *meant* to hit.[11]

Likewise Kazak et al. (1997) presented autistic children with trials on which either they, or a third party, were allowed to look inside a box, or were not allowed to look inside a box. They were then asked whether they or the third party knew what was in the box, or were just guessing. The autistic children got many more of these questions wrong than did control groups. And importantly for our purposes, there was no advantage for answers to questions about the child's own knowledge over answers to questions about the knowledge of the third party. This result is especially striking because the children *could* have answered the self-knowledge version of the question merely by asking themselves the first-order question, "What is in the box?", without needing to engage in metacognitive processes at all (except when transforming the result into a metacognitive answer to the experimenter's question).

I conclude that data from people with autism provide no support for the view that metacognition can remain intact in the absence of mindreading. On the contrary, the evidence suggests that if mindreading is damaged, then so too will be metacognition. Now admittedly, this by itself is just as consistent with model 2 ("one mechanism, two modes of access") as with model 4 ("mindreading is prior"). But our discussion in section 9 failed to find the alleged evidence that might speak in favor of the former (i.e., individuals in whom mindreading is intact but metacognitive access is blocked). And we have discussed a variety of other forms of evidence that support the latter.

11. Conclusion

This target article has evaluated four different accounts of the relationship between mindreading and metacognition, three of which endorse the existence of introspection for attitudes, whereas the fourth denies it. Since we know that people have the illusion of introspecting even when they demonstrably aren't doing so, and since design considerations suggest that the mindreading faculty would picture the mind as having introspective access to itself, I have argued that no weight should be placed on the introspective intuition. In which case the "mindreading is prior" account should be accepted by default, as the simplest of the four possibilities. In addition, I have argued that various predictions made by the three accounts that endorse introspection for attitudes are not borne out by the data. In contrast, the central prediction of the "mindreading is prior" account is confirmed: This is that subjects should be caused to misattribute attitudes to themselves by misleading sensory or behavioral data. Although an introspection theorist can attempt to save this data post hoc, such attempts are less than convincing. Hence, the "mindreading is prior" account is, overall, the best supported of the four alternatives.

ACKNOWLEDGMENT

I am grateful to the following for their helpful comments on a previous draft of this article: José Bermúdez, Paul Bloom, Daniel Dennett, Shaun Nichols, Rebecca Saxe, and an anonymous reviewer. In addition, I am grateful to the students in my graduate seminar on this topic, who critiqued my work and helped me to think through the issues: Mark Engleson, Marianna Ganapini, Yu Izumi, David McElhoes, Christine Ng, Elizabeth Picciuto, Vincent Picciuto, Yashar Saghai, Elizabeth Schechter, and Sungwon Woo, with special thanks to Mark Engelbert and Brendan Ritchie.

NOTES

1. One might wonder why the dedicated input channels between the various perceptual systems and the metarepresentational faculty couldn't be damaged while leaving those systems

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

themselves intact. The answer is that there are no such channels. Rather, the attended outputs of perception are globally broadcast to all conceptual systems, including the metarepresentational faculty inter alia. See section 2 for some discussion and references.

2. All of these authors endorse broadly "theory-theory" accounts of mindreading. A very different kind of "mindreading is prior" account is defended by Gordon (1986; 1996), who develops a form of simulation theory that denies any need for introspection. But this account makes both mindreading and metacognition dependent upon the acquisition of natural language. Likewise, Dennett (1991) is a sort of theory-theorist who denies introspection for attitudes, but he, too, appears to make our knowledge of our own mental states dependent upon their expression in language. Discussion of these issues would take us too far afield. For present purposes I assume, as seems plausible, that basic capacities for both mindreading and metacognition are independent of our capacity for natural language.

3. Note that for this reason Nichols and Stich's (2003) introduction of a separate perception-monitoring mechanism is wholly unnecessary. Since the mindreading system would need to have access to the agent's own perceptual states in order to do its work, there is simply no need for a distinct system to monitor and self-attribute those states.

4. In allowing that perceptual *judgments* are introspectable, I don't mean to imply that perceptually based *beliefs* are likewise introspectable. On the contrary, once formed and stored, the only way that those beliefs can be consciously accessed is via their expression in visual imagery (in the form of an episodic memory, perhaps) or in inner speech. But such events, although introspectable, will need to be interpreted to extract the information that they are, indeed, expressive of belief (as opposed, for example, to supposition or mere idle fantasy). See section 2.1 for further discussion.

5. An alternative account to the one sketched here is outlined by Wilson (2002), who suggests that the introspective assumption may make it easier for subjects to engage in various kinds of adaptive self-deception, helping them build and maintain a positive self-image. In fact, *both* accounts might be true.

6. We also know that in other domains – such as physics – the unconscious theories that guide behavior often make false, but simplifying, assumptions. See, for example, McCloskey (1983).

7. This isn't quite accurate. For, to the extent that apes, for example, do have limited mindreading abilities (e.g., in respect of perception and goal-directed action), to that extent one might expect to find metacognitive processes also. At any rate, this is what a "mindreading is prior" account would predict.

8. *Sometimes* a System 2 utterance *does* express an underlying System 1 judgment with the same content, no doubt. But in such a case it is all the clearer that the utterance in question isn't *itself* a judgment. Nor does the expressibility of judgments in speech provide any reason for believing in introspection, as we saw in section 2.1.

9. Similar claims are made by Bayne and Pacherie (2007). They argue against an interpretative account of self-awareness of the sort defended here, preferring what they call a "comparator-based" account. But I think they mis-characterize the models of normal action-monitoring that they discuss. Properly understood, those models lend no support for the claim that metacognition is damaged in schizophrenia. See the paragraphs that follow.

10. The claim that we have introspective access to our own motor intentions seems also to underlie the idea that "mirror neurons" might play an important role in the development of mindreading (Gallese & Goldman 1998). For what would be the use, for purposes of social understanding, of an activation of one's own motor system in response to an observation of the action of another, unless one could acquire metacognitive access to the motor plan in question? (For a variety of criticisms of this account of the mirror neuron system, see Csibra [2007] and Southgate et al. [2008].)

11. Russell and Hill (2001), however, were unable to replicate these results. This is probably because their population of autistic children, although of lower average age, had higher average verbal IQs, suggesting that their autism was much less severe. Since most researchers think that intention-reading is among the easiest of mindreading tasks, one might predict that only very young or more severely disabled individuals with autism would be likely to fail at it.

Open Peer Commentary

What puts the "meta" in metacognition?

doi:10.1017/S0140525X09000557

Michael L. Anderson[a,b] and Don Perlis[b,c]

[a]*Department of Psychology, Franklin & Marshall College, Lancaster, PA 17604;* [b]*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742;* [c]*Department of Computer Science, University of Maryland, College Park, MD 20742.*

michael.anderson@fandm.edu **http://www.agcognition.org**
perlis@cs.umd.edu **http://www.activelogic.org**

Abstract: This commentary suggests an alternate definition for metacognition, as well as an alternate basis for the "aboutness" relation in representation. These together open the way for an understanding of mindreading that is significantly different from the one advocated by Carruthers.

Carruthers suggests that cognitive scientists are confused about the meaning of "metacognition," citing our work as an illustrative example. In fact, we follow a standard definition of the term, adopted from Nelson and Narens (1990). (This particular formulation appears in Anderson & Oates [2007], but the definition is in widespread use. See, e.g., Dunlosky 2004; Dunlosky & Bjork 2008; Dunlosky & Metcalfe 2009; Metcalfe 1993; Metcalfe & Shimamura 1994.) The definition runs as follows:

> Imagine two components *X* and *Y* (where *X* and *Y* could be the same), related in such a way that state information flows from *Y* to *X*, and control information flows from *X* to *Y*. Component *X* is in a monitoring and control relationship with *Y*, and when *Y* is a cognitive component, we call this relationship metacognitive monitoring and control.

This offers an information-theoretic characterization of metacognition that is neutral regarding the form that information takes, or the processing it undergoes. Thus, it is quite incorrect to say that cognitive scientists use the term "in two quite distinct ways, often without noticing the difference" (target article, sect. 5.1, para. 2). We use the term consistently in a way that leaves open the various ways in which such a relationship could be implemented. We are not confused about the difference between systems that involve "metarepresentations of [its] own first-order cognitive processes as such" (sect. 5.1, para. 2) and those that don't; rather, this distinction is not relevant to the definition of metacognition.

In fact, *some* of the processes in the systems we implement are indeed metacognitive in Carruthers' more restricted sense. To take just one example, mentioned by Carruthers: If an active logic system notices the presence of both P and $\neg P$ in its knowledge base (KB), it will assert $Contra(P, \neg P, t)$.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

That is a statement *about* – a metarepresentation of – the state of the KB at time *t* (i.e., that it contained that contradiction). Our systems can reason about this fact with that metarepresentation, and consequently take various control steps, the simplest of which is to refrain from using these premises in further deduction (Anderson & Perlis 2005a). But other processes in active logic systems, and other of our metacognitive systems, effect such monitoring and control without explicit metarepresentations of this sort (see, e.g., Anderson et al. 2006).

Of course, Carruthers is free to define his terms and circumscribe his interests as best serves his argument, and if this were merely a terminological dispute, we would not be submitting a commentary. But there is a more substantive point in the background, which potentially affects Carruthers' overall proposal. Carruthers writes: "Generally the term is used, as it has been throughout this article, to mean cognition *about* one's own cognition. Metacognition, in this sense, is inherently higher-order, involving metarepresentations of one's own first-order cognitive processes as such" (sect. 5.1, para. 2, emphasis in original). The implication seems to be that for something to be *about* another requires a higher-order metarepresentation. But we would like to suggest that this associates *higher-order*-ness with *meta*-ness and *about*ness (if we can be forgiven the neologisms) in a way that is not necessary.

First, it is not clear that aboutness requires higher-order-ness. Surely a representation or a process can be about another without being at a different level, or in a different representational language. Indeed, can't a process (or representation) be about itself? (See, e.g., Perlis 1985; 1988; 1997; 2000; Perlis & Subrahmanian 1994.) It is a common bias, perhaps stemming from Tarski, that there must be a hierarchy of meta-languages, each standing back from the one it refers to. But Tarski adopted that approach to avoid technical difficulties in formal logic; it is not necessary a priori.

Second, it is not clear that meta-ness requires higher-order-ness. In related writings, we have suggested that representation requires only the following: tokens, whatever their form/content, that can be used to guide actions with respect to certain targets (Anderson & Perlis 2005b; Anderson & Rosenberg 2008). On these accounts, the information being used and manipulated during cognition is representational just in case it is used to guide behavior with respect to targets in various circumstances. Likewise, a metacognitive monitoring and control process represents a cognitive process, just in case it allows the metacognitive component to guide actions with respect to the cognitive process. Such monitoring and control is indeed (we maintain) cognition *about* cognition – is thus *meta*cognition – without having to be/utilize higher-order representations of cognition as such.

As should be clear from the preceding, we have a somewhat different understanding of what the representational aboutness relation requires. This most definitely applies to self-representation as well (Anderson & Perlis 2005b), although it is perhaps worth noting that the account of self-awareness we develop in the cited paper is – despite differences in the fundamental criteria for aboutness – nevertheless compatible with the "mindreading is prior" framework that Carruthers advocates.

So why might all of this matter to Carruthers? Because of Carruthers' understanding of what aboutness requires, he is driven to adopt a higher-order, meta-representational account of what having certain thoughts about another's thoughts ("mindreading") requires. In contrast, the less restrictive option offered by us opens the door for a broader range of theories of what our responsiveness to the mental states of others requires. This would include, for instance, Shaun Gallagher's interesting, and interestingly different, interaction-based account of understanding self and others (Gallagher 2004; 2005). It would have been useful and instructive to see how this rather broader portrayal of the competing possibilities might have affected Carruthers' argument, discussion, and conclusions.

Is feeling pain just mindreading? Our mind-brain constructs realistic knowledge of ourselves

doi:10.1017/S0140525X09000569

Bernard J. Baars

The Neurosciences Institute, San Diego, CA 92121.

baarsbj@gmail.com

http://bernardbaars.pbwiki.com

Abstract: Carruthers claims that "our knowledge of our own attitudes results from turning our mindreading capacities upon ourselves" (target article, Abstract). This may be true in many cases. But like other constructivist claims, it fails to explain occasions when constructed knowledge is *accurate*, like a well-supported scientific theory. People can know their surrounding world and to some extent themselves. Accurate self-knowledge is firmly established for both somatosensory and social pain.

Brain imaging studies show that social pain (like social rejection, embarrassment, and guilt) activates brain regions characteristic of painful *bodily* experiences. The brain regions that are activated by both evoked social and physical pain include the anterior cingulate cortex, the right prefrontal lobe, the insula, amygdala, and somatosensory cortex. Even deep brain structures, such as the brainstem periaqueductal gray (PAG), are known to be evoked by mother–infant separation, marked by intense and repeated distress cries. These functions are highly conserved among mammals and, perhaps, birds (Eisenberger & Lieberman 2004; Nelson & Panksepp 1998).

This evidence contradicts Carruthers' hypothesis that we learn about ourselves by turning our social mindreading capacities upon ourselves. No doubt we do learn about ourselves based upon what we have learned about others. After all, we constantly transfer knowledge between different domains of reference. However, it is simply not the case that *all* of our introspective self-knowledge is of this kind. Children acquire "theory of mind" abilities in about the fourth year of life. But long before that time we can observe, pain and pleasure perception, the distress of abandonment, anticipatory fear and joy, and a wide spectrum of social and imaginary emotional experiences.

Carruthers could maintain that such emotional experiences are not true cases of "metacognition" and "introspection." It is possible to define such terms in very limited ways, but there is no doubt that emotional feelings express propositional attitudes: They are *about* something, namely the well-being of the self. Thus, hunger, thirst, air-hunger, social distress, fear of rejection by the mother, peer envy, and numerous other infant emotions are by no means simple "reflexes." They are socially contingent, though not explicitly deliberated, reactions to real-world events that are critical to the infant's survival. This crucial self-related information has extraordinary breadth of conservation among mammals, suggesting an evolutionary history of some 200 million years (Baars 2005).

Pain is not the only kind of introspective experience humans have with minimal social input, but it is perhaps the most compelling. Metacognitive self-report ("introspection") has been used for two centuries in psychophysics. It is a well-established methodology that converges extremely well with other empirical evidence, such as brain recording methods (Baars & Gage 2007).

Science is a constructive enterprise, but it is tightly constrained by evidence. That is why, like other human activities such as farming and tax accounting, it is not merely constructed, but also bound by considerations of accuracy and predictability.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

That is true for humans, but it is equally true for animals, who must survive real-world challenges in environments in which errors lead to extinction. Brain evolution is not separate from the ability to observe and know the real world. On the contrary, when we are given truthful feedback about the world, humans and other animals become quite reality-based. There is no contradiction between constructivism and realism.

How "weak" mindreaders inherited the earth

doi:10.1017/S0140525X09000570

Cameron Buckner,[a] Adam Shriver,[b] Stephen Crowley,[c] and Colin Allen[d]

[a]*Department of Philosophy, Indiana University, Bloomington, IN 47405-7005;* [b]*Philosophy-Neuroscience-Psychology Department, Washington University in St. Louis, St. Louis, MO 63130;* [c]*Department of Philosophy, Boise State University, Boise, ID 83725-1550;* [d]*Department of History and Philosophy of Science, Indiana University, Bloomington, IN 47405.*

cbuckner@indiana.edu
http://www.indiana.edu/~phil/GraduateBrochure/IndividualPages/cameronbuckner.htm **ajshrive@artsci.wustl.edu**
http://artsci.wustl.edu/~philos/people/index.php?position_id=3&person_id=60&status=1
stephencrowley@boisestate.edu
http://philosophy.boisestate.edu/Faculty/faculty.htm
colallen@indiana.edu
http://mypage.iu.edu/~colallen/

Abstract: Carruthers argues that an integrated faculty of metarepresentation evolved for mindreading and was later exapted for metacognition. A more consistent application of his approach would regard metarepresentation in mindreading with the same skeptical rigor, concluding that the "faculty" may have been entirely exapted. Given this result, the usefulness of Carruthers' line-drawing exercise is called into question.

Carruthers' recent work on metacognition in the target article (and in Carruthers 2008b) can be seen as an extended exercise in "debunking" metarepresentational interpretations of the results of experiments performed on nonhuman animals. The debunking approach operates by distinguishing "weak" metacognition, which depends only on first-order mechanisms, from "genuine" metacognition, which deploys metarepresentations. Shaun Gallagher (2001; 2004; with similar proposals explored by Hutto 2004; 2008) has been on a similar debunking mission with respect to metarepresentation in human mindreading abilities. Gallagher's position stands in an area of conceptual space unmapped by Carruthers' four models, which all presuppose that an integrated, metarepresentational faculty is the key to mindreading. Gallagher argues that most of our mindreading abilities can be reduced to a weakly integrated swarm of first-order mechanisms, including face recognition and an ability to quickly map a facial expression to the appropriate emotional response, a perceptual bias towards organic versus inorganic movement, an automated capacity for imitation and proprioceptive sense of others' movements (through the mirror neuron system), an ability to track the gaze of others, and a bias towards triadic gaze (I-you-target). Notably, autistic individuals have deficiencies throughout the swarm.

Someone pushing a "metarepresentation was wholly exapted" proposal might argue as follows: Interpretative propositional attitude ascription is a very recent development, likely an exaptation derived from linguistic abilities and general-purpose concept-learning resources. Primate ancestors in social competition almost never needed to think about others not within perceptual range; in the absence of language which could be used to raise questions and consider plans concerning spatially or temporally absent individuals, there would have been little opportunity to demonstrate third-person mindreading prowess. After developing languages with metarepresentational resources, our ancestors' endowment with the swarm would have left them well placed to acquire metarepresentational mindreading and metacognition through general learning. While such abilities were likely favored by cultural evolution in comparatively recent history, it is not clear that any further orders to genetic evolution needed to be placed or filled. Evolutionary "just so" stories come cheap; if Carruthers wants to make a strong case that the faculty evolved in response to social pressures (instead of just excellence with the swarm and/or other general aspects of cognition thought to be required for Machiavellian Intelligence, such as attention, executive control, and working memory), he needs further argument.

Two issues must be overcome for the swarm proposal to be considered a serious alternative. First, the concurrent appearance of success on verbal first- and third-person false-belief tasks must be explained. Here, we point the reader to Chapter 9 of Stenning and Van Lambalgen (2008), which makes a strong case that the logic of both tasks requires a kind of conditional reasoning which does not develop until around age 4 and is also affected by autism (and see also Perner et al. [2007] for a related account). Second, there is the work on implicit false-belief tasks with prelinguistic infants (Onishi & Baillargeon 2005). These findings are both intriguing and perplexing (consider, for example, that the infants' "implicit mastery" at 15 months is undetectable at 2.5 years), and the empirical jury is still out as to whether the evidence of preferential looking towards the correct location can support the weight of the metarepresentational conclusions which have been placed on it (see Perner & Ruffman 2005; Ruffman & Perner 2005). The infants' preferential looking can be explained if they quickly learn an actor-object-location binding and register novelty when the agent looks elsewhere. More recent studies (e.g., Surian et al. 2007) claiming to rule out alternatives to the metarepresentational explanation have produced findings that are ambiguous at best (Perner et al. 2007).

One might concede that the mechanism generating the gaze bias in infants is not itself metarepresentational, but nevertheless hold that it evolved because it enabled its possessors to develop metarepresentation – likely wielding a poverty of the stimulus (PoS) argument to the effect that even with language, metarepresentational mindreading does not come for free. We suggest that such reasoning no longer carries the weight it once did. Recent work on neural network modeling of the hippocampus, which highlights its ability to quickly discover abstract, informationally efficient bindings of stimulus patterns (especially when fed neutral cues like words – e.g., see Gluck & Myers 2001; Gluck et al. 2008) dulls the PoS sword. Finally, even if the PoS argument is accepted, there remains a huge leap to the conclusion that the bias evolved *because of its ability to bootstrap metarepresentation* – and not for something simpler.

In light of the swarm alternative, the usefulness of Carruthers' distinction between "weak" and "genuine" forms of mindreading and metacognition becomes questionable. Our overarching worry is that Carruthers' emphasis on a single faculty of metarepresentation, combined with his acknowledgment of the rich heritage of cognitive abilities shared between humans and animals, leaves the faculty almost epiphenomenal in human cognition (except, perhaps, for Machiavelli himself) – a position that Carruthers has previously been driven to adopt with respect to his account of phenomenal consciousness (Carruthers 2005; see also Shriver & Allen 2005). An alternative approach might be to tone down the deflationary invocation of first-order mechanisms, and focus instead on what creatures endowed with a swarm of weakly integrated mechanisms can do and learn. Once we abandon the assumption that mindreading is centralized in a single metarepresentational faculty, we can investigate whether something like Gallagher's swarm could implement various degrees of competence in reacting adaptively to the mental states of others. This perspective focuses us on the flexibility and adaptive significance of the evolved mechanisms which

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

constitute the swarms, for a wide range of organisms in a variety of social environments (including humans in theirs). These suggestions are in the spirit of Dennett (1983), who advocated the usefulness of metarepresentational hypotheses in devising new experiments, accepting from the beginning that animals and humans will "pass some higher-order tests and fail others" (p. 349). Ultimately, we think that the questions Carruthers raises about the relationship between self-regarding and other-regarding capacities are interesting and should be pursued; and they *can* be pursued without engaging in the line-drawing exercise which de-emphasizes the significance of good comparative work for understanding human cognition.

ACKNOWLEDGMENT

We thank Jonathan Weinberg for his extensive comments on earlier versions of this commentary.

Cognitive science at fifty

doi:10.1017/S0140525X09000582

A. Charles Catania

Department of Psychology, University of Maryland, Baltimore County (UMBC), Baltimore, MD 21250.

catania@umbc.edu

http://www.umbc.edu/psyc/personal/catania/catanias.html

Abstract: Fifty years or so after the cognitive revolution, some cognitive accounts seem to be converging on treatments of how we come to know about ourselves and others that have much in common with behavior analytic accounts. Among the factors that keep the accounts separate is that behavioral accounts take a much broader view of what counts as behavior.

Roughly half a century has passed since the cognitive revolution declared behaviorism dead and promised solutions to long-standing problems of philosophy and psychology. Carruthers provides an opportunity to assess the progress that has taken place. Mind remains central in his account, and its hierarchical structure is illustrated in the pivotal roles of metarepresentations and metacognitions. In place of behavior and events in the world, the action takes place in the dynamics of their surrogates, such as perceptions and intentions and beliefs and concepts and attitudes, none of which lend themselves to measurement in the units of the physical or biological sciences. Most of the entities in Carruthers' account existed in the vocabularies of the mid-1950s, though typically more closely anchored to their origins in colloquial talk, which since then has sometimes been called folk psychology.

What has most obviously changed are the linkages among the mentalistic terms. Carruthers deals with the particular priorities of mindreading and metacognition. Are they independent mechanisms or a single mechanism with two modes of access? Is one a prerequisite for the other? Carruthers concludes that metacognition is grounded in mindreading. If one argues that judgments about oneself must be distinguished from judgments about others, his conclusion is sound. But this conclusion is one that a variety of behaviorism reached long before the advent of the cognitive revolution. In his "Behaviorism at 50," Skinner (1963) recounted the history of Watsonian methodological behaviorism in the early decades of the twentieth century and its rejection of introspection (see also Catania 1993), but he also noted the unnecessary constraints that Watson's account had imposed on theory.

Skinner's later radical behaviorism rejected the Watsonian constraints and extended his approach to the origins of the language of private events. As a contribution to a symposium organized by his advisor, E. G. Boring, Skinner (1945) made explicit his interest in "Boring from Within." The 1945 paper, "The Operational Analysis of Psychological Terms," was a renunciation of operationism, but, more important, it provided an account of how a vocabulary of private events (feelings, emotions, etc.) could be created even though those who taught the words and maintained consistencies of usage had access only to shared public accompaniments of those private events.[1] Given these origins of the private or introspective language, Skinner's resolution of the issue in terms of the public practices of the verbal community is the only feasible way of dealing with the problem that Carruthers has so aptly described in terms of his mindreading system, which never has access to what others are imagining or feeling. To the extent that it does have access to what one feels or imagines oneself, one can speak of those events only in a vocabulary that is anchored in public correlates. Carruthers' point that instances of self-attributed unsymbolized thought occur in circumstances in which a third party might have made the same attribution is perfectly consistent with this argument.

The irony, then, is that with respect to introspection, judgments about the behavior of others (mindreading) and judgments about one's own behavior (metacognition), Carruthers has reached conclusions that are consistent with Skinner's. One can guess that he took so long only because of the complexity of the terms that entered into his account. Skinner's account is far more parsimonious. Skinner does not begin with something called discriminating and follow it with differential responding; the differential responding is itself the discriminating. He does not say that perceiving and sensing and thinking are something different from behaving; they are kinds of behavior, defined not by whether they involve movement but rather by whether they are involved in contingent relations with environmental events (for this reason, Carruthers notwithstanding, a lot of behavior goes on even when one is sitting quiet and motionless, and one has just as much access to this behavior as to that of standing or walking). There is no more need to appeal to seeing and hearing as prerequisite concepts than there is to say that we cannot sit or stand or walk without concepts of sitting or standing or walking; these are all names for things we do. To invoke them as explanations does not serve our theories well.

Carruthers' account also converges on other concepts that have been elaborated by Skinner. For example, his System 1 and System 2 have features that are closely paralleled by what Skinner (1969) respectively called rule-governed and contingency-shaped behavior, and Carruthers is surely on the right track in saying that speech is an action that does not begin with metacognitive representations of thought (a more detailed account is beyond the scope of this commentary, but see Catania 2006, Chs. 14 and 15). Furthermore, in considering the different environmental contingencies that operate on verbal and nonverbal classes of behavior, the behavioral account has no trouble dealing with the various confabulations that Carruthers has surveyed. Just as speech errors can tell us a lot about language structure, so confabulations may tell us a lot about the nature of our judgments about ourselves and others.

It is good to see cognitive science at last converging on conclusions that had once been reached in behavioral accounts. If that were the only point, this commentary would serve little but a historical purpose. But there is extensive behavior analytic research relevant to these issues (in particular, see Wixted & Gaitan 2002), and some of it may prove useful to those of any theoretical orientation. Of course, it would be not at all surprising if the suggestions here are not well received. That likelihood is enhanced by the fact that this has been a necessarily brief and superficial presentation of the behavioral case. But the literature is there, so perhaps a few will check it out.

NOTE

1. Two articles by B. F. Skinner cited in this commentary (Skinner 1945; 1969) were reprinted in *Behavioral and Brain Sciences* (Vol. 7, December 1984).

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Metacognition is prior

doi:10.1017/S0140525X09000594

Justin J. Couchman,[a] Mariana V. C. Coutinho,[a] Michael J. Beran,[b] and J. David Smith[a]

[a]*Department of Psychology, University at Buffalo, The State University of New York, Buffalo, NY 14260;* [b]*Language Research Center, Georgia State University, Atlanta, GA 30303.*

jjc38@buffalo.edu **mvc5@buffalo.edu**
mjberan@yahoo.com **psysmith@buffalo.edu**

Abstract: We agree with Carruthers that evidence for metacognition in species lacking mindreading provides dramatic evidence in favor of the *metacognition-is-prior* account and against the *mindreading-is-prior* account. We discuss this existing evidence and explain why an evolutionary perspective favors the former account and poses serious problems for the latter account.

Carruthers acknowledges that evidence for metacognition in species lacking mindreading would provide dramatic evidence for the *metacognition-is-prior* view and against the *mindreading-is-prior* view, and he asserts that the existing evidence can be explained using a first-order system of belief and desire *strengths* (target article, sect. 5.2; see also Carruthers 2008b). We evaluated similar response strategies using formal modeling (Smith et al. 2008) and found indeed that some animal metacognition findings could be explained using first-order strategies. Yet Carruthers' use here of the field's earliest paradigms and oldest data to make his argument is unfortunately selective. More recent paradigms often do not support his first-order argument and description.

Smith et al. (2006) dissociated monkeys' uncertainty responding from any reinforcement and stimulus cues that could have organized Carruthers' gradients of first-order beliefs and response tendencies. It was clear in that study that monkeys' uncertainty-response strategies were adjudicated cognitively and decisionally, not using first-order cues. They followed the animal's subjective decisional construal of the task. Couchman et al. (submitted) extended this dissociation to situations of broader task transfer in which animals had to establish functional regions of judged difficulty and uncertainty even when forced to self-organize their task performance.

Recent cross-species research on uncertainty monitoring also speaks against first-order interpretations of uncertainty-monitoring performances. Beran et al. (in press) gave capuchin monkeys a Sparse-Uncertainty-Dense task that was matched to a Sparse-Middle-Dense task. Capuchins used the middle (first-order) response easily and naturally. They almost never used the uncertainty response, despite having the reinforcement history needed to do so. Likewise, elegant research by Shettleworth and her colleagues (Inman & Shettleworth 1999) has shown that pigeons also do not express an uncertainty-responding capability, even when there are strong first-order reasons for them to do so. It is an important implication from these cross-species results that the organizing psychology underlying uncertainty responding is not first-order, because adept first-order animals such as capuchins and pigeons cannot find and use that psychology.

In other writings, Carruthers (2008b) also acknowledges that first-order beliefs and desires will not explain the wide-ranging empirical findings of uncertainty monitoring and information seeking by animals. He devises a secondary mental construct to explain why an animal uses the uncertainty response in too-close-to-call situations. He suggests that some species have a gate-keeping "mechanism . . . 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" (p. 66).

The gatekeeper mechanism operates on first-order cognition's outputs to assess their ability to produce a correct response. It meets the definition of a second-order controlled cognitive process. It produces a qualitative change in behavior and cognitive strategy (information seeking, uncertainty responses, etc.). It typifies the metacognitive utility that all theorists have envisioned. Even in Carruthers' own description of animals' cognitive self-regulation, it seems, metacognition is prior.

Another analytic problem in the target article concerns the different standard of evidence that is applied to studies of animal metacognition and studies of animal mindreading. It seems highly unlikely, and it goes completely undefended in the target article (sect. 5, para. 2) that all the metacognition paradigms fall prey to behaviorist explanations, but that all the mindreading paradigms are veridical. They clearly are not (Heyes 1998).

Carruthers makes a valid suggestion that, if metacognition is prior, one should be able to explore the evolutionary pressures that produced a free-standing metacognitive utility. Fortunately, James (1890/1952), Dewey (1934/1980), Tolman (1938), and many others have provided this evolutionary narrative (see also Smith et al. 2003). Animals often encounter doubtful and uncertain situations in which their habitual stimulus-response associations do not clearly indicate a safe and adaptive response. They would benefit enormously in those situations from having an online cognitive utility that will let them assemble the relevant facts and recollections and choose an adaptive course of action. Metacognition provides exactly this utility.

It is also a remarkable phylogenetic fact that there appear to be no species that show mindreading ability but fail to show metacognitive ability. This could be used to support more than one of the possibilities discussed in the target article. However, it clearly supports least of all the *mindreading-is-prior* account.

Finally, we believe that an evolutionary perspective on this issue raises a serious problem for the *mindreading-is-prior* account. The author's account may, in principle, explain the development of metacognition ontogenetically, especially if one assumes a parent is constantly informing you of the intentions of others. Your mother may tell you, "Johnny wants a cookie" while you see Johnny reaching for the cookie jar, and the next time you find yourself reaching for the cookie jar, you may well apply "wants a cookie" to yourself. This works only because humans communicate their knowledge of concepts and intentions from one generation to the next.

The first mindreading animal would have no basis for which to make an attribution of a mental state. How would it be possible or beneficial to attribute "wants a cookie" to Johnny, if the attributer has no known experience with "wanting," no understanding of "what it is like to want" and no idea that it has ever "wanted"? The *mindreading-is-prior* account must explain how, from nothing but observed physical behavior, and with no reason to ever attribute anything but cause-and-effect mechanical processes, animals came to attribute subjective belief and desire states to others. This would be equivalent to knowing there is something "that it is like" to be a bat (Nagel 1974) prior to knowing that there is anything "that it is like" to be you!

Indeed, exactly the opposite seems to be true. We have great access to and a rich understanding of our own mental states and only a very limited understanding of the mental states of others. We first knew what it was like to know, and then assumed that others might be having an analogous experience. This process of extending mental concepts outward is surely a more plausible and tractable evolutionary narrative. Within that narrative, metacognition is prior.

Introspection, confabulation, and dual-process theory

doi:10.1017/S0140525X09000600

Jonathan St. B. T. Evans

Centre for Thinking and Language, School of Psychology, University of Plymouth, Plymouth PL4 8AA, United Kingdom.

j.evans@plymouth.ac.uk

Abstract: This excellent target article helps to resolve a problem for dual-process theories of higher cognition. Theorists posit two systems, one of

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

which appears to be conscious and volitional. It seems to control some behaviours but to confabulate explanations for others. I argue that this system is only conscious in an illusory sense and that all self-explanations are confabulatory, as Carruthers suggests.

I have long held (Evans 1980) that while we can introspect on our mental experiences, we have no access to the processes which underlie our behaviour, and I have equally long held the view (Evans 1989; Wason & Evans 1975) that strategy reports frequently reflect confabulations. Crossing the disciplinary divide with philosophy, this accords well with Carruthers' claims that (a) we have no introspective access to our propositional attitudes and (b) that we may have an illusion of conscious control resulting from applying our "mindreading" abilities to ourselves as well as others. Although Carruthers seeks to reconcile his massively modular view of the mind with dual-process theories of reasoning (see also Carruthers 2006), it is not clear how many tenets of the standard theory he would accept. I therefore examine his "mindreading is prior" argument with respect to this standard dual-process approach.

Dual-process theorists propose that humans have two distinct forms of cognitive processing: one fast, automatic, and high capacity (Type 1), and another slow, controlled, and low capacity (Type 2) (Evans 2008). It is commonly assumed that these two modes of thought reflect two distinct forms of knowledge: implicit and explicit (Carruthers does not appear to endorse this distinction). Implicit knowledge may be encapsulated in cognitive modules or acquired from associative or procedural learning. Explicit knowledge has some propositional format and can be "called to mind." Neuroscientific evidence strongly supports the existence of dissociable implicit and explicit memory systems (Eichenbaum & Cohen 2001). Intuitive and reflective judgments are assumed to reflect access to these two forms of knowledge and to comprise two distinct cognitive systems. This can explain, for example, why people seem to possess implicit attitudes and stereotypes, which may conflict and compete with their explicit social attitudes (Smith & DeCoster 2000).

There is, however, a point of discomfort within dual-process theory to which the current target article is highly relevant. Since theorists claim that there are two systems with functional control of behaviour, System 2 – the "conscious" one – cannot be epiphenomenal. On the other hand, as Carruthers correctly states, the evidence that people frequently confabulate explanations for their behaviour is overwhelming. Evans and Over (1996) were clearly struggling with this issue, when they stated (p. 160) that "we do not regard explicit thinking as simply serving to rationalise behaviour, and believe that decisions and actions can result from explicit processes." Does this mean that Type 2 processes sometimes *really* control a response and at other times confabulates an explanation for a Type 1 response? Could we not instead argue that *all* strategy reports are self-interpretations in just the same way as Carruthers argues for propositional attitudes?

The problem as I see it (Evans 2008; 2009) is that it is a mistake to use consciousness in the definition of the distinction between Systems 1 and 2. A more satisfactory definition is that the latter requires access to central working memory, whereas the former does not. For sure, this implies that *something* about a Type 2 process is conscious, as the contents of working memory tend to reflect in conscious experiences (Andrade 2001). However, it can still be the case that (a) most of the workings of System 2 are unconscious – for example, the processes that direct our current locus of attention, and those that retrieve memories relevant to the current context, and (b) that we lack introspective access to the nature of System 2 processing. (Note that [b] may be at odds with Carruthers' claim that we do have a conscious mind.) Introspection provides no access to Type 1 processes, that either (preconsciously) pass information into working memory for further processing, or by-pass it altogether (Evans 2009). Because System 2 uses working memory, something about it can be introspected – the locus of our attention, for example, or relevant perceptual and emotional experiences. These will be used as inputs to the self-interpretation process that Carruthers discusses, together with contextual knowledge and any current goals we are pursuing. Because we generally have good theories of our own behaviour, we can often produce veridical reports. On other occasions, we confabulate, but I suggest that mental processes in either case are *exactly the same*.

I suggest that it is unhelpful to describe dual-process theories as contrasting conscious and nonconscious processing, and that the theory in no ways rests upon either introspective access or any notion of conscious control. Dual-process research methods depend instead on the assumptions that (a) Type 1 and Type 2 processing are qualitatively and measurably distinct in their characteristics and outputs, and (b) that the balance between the two forms of processing is related to a number of well-defined variables. These include cognitive ability, motivation, time available for reasoning, and the presence of competing or distracting tasks. However, we still need some account of why confabulation is common and why we do *feel* as though we have conscious control of our behaviour. This is where I find Carruthers' "mindreading is prior" argument helpful. If our brains have an in-built folk psychology for understanding the minds of others, why would the same system not interpret our own behaviour?

In summary, Carruthers' account helps to resolve a problem for standard dual-process accounts of higher cognition. People, he argues, are conscious of perceptual and quasi-perceptual experiences, the latter being formed by mental rehearsal that involves inner speech and/or imagery. It is precisely such processes – that involve working memory – that cause people to confabulate. People do *not* confabulate accounts of Type 1 processes such as those underlying recognising a face, or understanding a sentence; they *do* confabulate explanations for Type 2 processes such as decision making. As Carruthers says, people will also confabulate on the basis of behavioural data, but they omit reference to relevant Type 1 processes when they do so. For example, people never refer to a number of well-documented perceptual, cognitive, and social biases in their verbal reports, for folk psychology knows nothing of such matters. This also explains our chronic tendency to overly attribute conscious reasons for actions in ourselves and in others.

What can we say about the inner experience of the young child?

doi:10.1017/S0140525X09000612

Charles Fernyhough

Institute of Advanced Study, Durham University, Durham DH1 3RL, United Kingdom.

c.p.fernyhough@durham.ac.uk

http://www.dur.ac.uk/c.p.fernyhough

Abstract: Inner experience is proposed as a basis for self-interpretation in both children and adults, but young children's inner experience may not be comparable to our own. I consider evidence on children's attribution of inner experience, experience sampling, and the development of inner speech, concluding that Carruthers' theory should predict a developmental lag between mindreading and metacognition.

Carruthers' "mindreading is prior" model holds that we gain knowledge of our own propositional attitudes through applying our mentalizing capacities to our own behavior and inner experience. In evaluating this claim, Carruthers considers the question of developmental asymmetry between self- and other-knowledge. Both the "two independent mechanisms" and the "metacognition is prior" views would predict that metacognition

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

should appear before mindreading. These two models do not, however, exhaust the possibilities for non-interpretative accounts of metacognition (specifically, they leave "one mechanism, two modes of access" as a remaining competitor account). This fact, together with recent evidence for very early mentalizing competence which, for methodological reasons, cannot be matched by data on early metacognition, means that such evidence cannot discriminate between Carruthers' interpretative account and a non-interpretative alternative.

There is one respect, however, in which a developmental asymmetry between mindreading and metacognition could have a bearing on Carruthers' model. Carruthers proposes that self-interpretation can proceed on the basis of information about overt behavior and physical circumstances, along with elements of inner experience such as inner speech, visual imagery, and feelings. He notes that there will be some instances where, behavioral data being lacking (as in the example of someone sitting quietly in their living room), self-interpretation will be based exclusively on information about inner experience.

The question, then, is what sort of information that could support self-interpretation is available to young children. Presumably, young children have the same possibilities for interpreting external information as adults do. But is their inner experience comparable? At least three sources of evidence lead us to scepticism on this point.

First, there are the findings of experimental research on children's understanding of inner experience. For example, consider the findings of Flavell and colleagues (e.g., Flavell et al. 1993; 2000) that preschool children frequently deny the presence of inner experience in individuals (including themselves) when it would be appropriate to attribute such experience. These results are usually interpreted as evidence that young children have only weak powers of introspection. But an alternative interpretation is that young children do not experience a stream of consciousness in the way that older children and adults do, and that this accounts for their weak understanding of the inner experience of others (Hurlburt & Schwitzgebel 2007, sect. 11.1.7.8; Meins et al. 2003).

A second line of evidence comes from the limited data from experience sampling in children. Descriptive Experience Sampling (DES; Hurlburt & Heavey 2006) involves careful interviewing around records made of inner experience shortly preceding a random electronic beep. Hurlburt describes an episode of DES with a nine-year-old boy who reported an image of a hole in his backyard containing some toys (Hurlburt & Schwitzgebel 2007, Box 5.8). When asked whether this image was an accurate description of his backyard, the participant replied that he had not yet had time to put all of the toys into the image. Hurlburt's conclusion from this and other instances of childhood experience sampling is that constructing a visual image is a skill that takes time to develop. Although much remains to be done in adapting experience sampling techniques for use with young children, the evidence currently available invites caution in making assumptions about young children's inner experience.

Third, there are theoretical reasons for not assuming that certain aspects of children's inner experience, particularly inner speech, are comparable to those of adults. The most fully developed theory of the development of inner speech is that of Vygotsky (1934/1987). In his theory, inner speech is the developmental outcome of the internalization of social speech via the transitional stage of private speech. Findings from the study of private speech suggest that its transformation into inner speech is unlikely to be complete until middle childhood (Winsler & Naglieri 2003). The view that there is a general shift towards verbal mediation of cognition in the early school-age years is supported by findings that phonological recoding of visually presented material in short-term memory tasks is linked to private speech use at this age (Al-Namlah et al. 2006). Speech that is still in the process of being internalized is likely to appear to the child's consciousness as something other than adult-like inner speech (Fernyhough et al. 2007). Further experimental research on the transition to verbal mediation, complemented by more developmentally sensitive experience sampling studies, should provide a clearer indication of when inner speech can be assumed to be present in young children.

There are reasons, then, for doubting that young children have access to the full range of inner experiences proposed by Carruthers to form the basis of self-interpretation. Because inner speech is one of the main sources of evidence supposed to feed into individuals' interpretations of their own propositional attitudes, the emergence of metacognition should be developmentally constrained by the emergence of inner speech. Other aspects of inner experience, such as visual imagery, are also likely to take time to develop. Given what we know about the timetable for the emergence of mindreading capacities (particularly the evidence for some mentalizing competence in the second year of life), Carruthers' theory should predict a developmental lag between mindreading and metacognition. An alternative for Carruthers would be to argue that behavioral and contextual evidence was sufficient for self-interpretation in young children, but then his account would be indistinguishable from that of Gopnik (1993).

Confabulation, confidence, and introspection

doi:10.1017/S0140525X09000624

Brian Fiala and Shaun Nichols

Department of Philosophy, University of Arizona, Tucson, AZ 85721.
fiala@email.arizona.edu sbn@email.arizona.edu
http://dingo.sbs.arizona.edu/~snichols/

Abstract: Carruthers' arguments depend on a tenuous interpretation of cases from the confabulation literature. Specifically, Carruthers maintains that cases of confabulation are "subjectively indistinguishable" from cases of alleged introspection. However, in typical cases of confabulation, the self-attributions are characterized by low confidence, in contrast to cases of alleged introspection.

What is confabulation? Carruthers' central argument hinges on this notion, so we need to get clear on what he has in mind. Carruthers doesn't present an explicit characterization, but the overall discussion suggests that the relevant confabulations are a class of first-person mental state attributions that are generated by an "interpretative" process, as opposed to an "introspective" process. By "interpretative," Carruthers means any process "that accesses information about the subject's current circumstances, or the subject's current or recent behavior, as well as any other information about the subject's current or recent mental life" (sect. 1.4, para. 3). This characterization seems too broad because introspection itself is supposed to be a process that accesses information about the subject's current mental life. But Carruthers means to count as interpretative only those processes that do not employ any "direct" access or any mechanism specifically dedicated to detecting one's current mental states.

On Carruthers' view, all attributions of propositional attitude events are, in fact, interpretative. So what is the relation between "confabulation" and "interpretation"? Here are several different possibilities:

1. Confabulations include all self-attributions that result from interpretation.

2. Confabulations include all *false* self-attributions that result from interpretation, and accurate interpretative self-ascriptions do not count as confabulatory.

3. Confabulations include only a proper subset of false self-attributions resulting from interpretation.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

We doubt that Carruthers has possibility 1 in mind, as this would mean that one is confabulating even when one quite consciously uses interpretative processes to discern one's past mental states. If Carruthers has option 3 in mind, then we need to know much more about what distinguishes the proper subset. As a result, we proceed on the assumption that possibility 2 captures what Carruthers has in mind.

Our experience with identifying our own current mental states is characteristically quick, accurate, and confident. By contrast, when it comes to attributing mental states to others, our attributions seem much slower, more accident prone, and unsure. This subjective difference is thought to provide prima facie evidence that we have (non-interpretative) introspective access to our own mental states. Carruthers attempts to defeat this prima facie consideration by proclaiming that confabulated reports are subjectively indistinguishable from cases of alleged introspection. People confabulate attributions of their own propositional attitude events "while being under the impression that they are introspecting" (sect. 6, para. 1). Thus, we have no reason to think that canonical cases of "introspection" differ from confabulation in this respect (i.e., that we are interpreting in the latter case but not the former). Carruthers goes on to argue that since there is no other positive reason to believe in the reality of introspection for the attitudes, the best explanation is that all self-attribution (confabulation and alleged introspection) is subserved by the same kinds of processes: that is, interpretative ones.

Carruthers' argument depends on the claim that people confabulate attributions of propositional attitudes while being under the impression that they are introspecting. But we are given no evidence that this has been systematically investigated. Certainly no one has ever asked participants in these cases whether they think they are introspecting or interpreting. Without some more direct evidence, Carruthers is not warranted in claiming that when people confabulate they are often "under the impression that they are introspecting."

A closer look at the confabulation cases gives further reason to doubt the argument. The evidence on confabulation cited by Carruthers is all anecdotal, but even the anecdotes are illuminating if one looks at the behavior a bit more closely. For we find that across many different paradigms in which people confabulate, the confabulations are not reported with a sense of "obviousness and immediacy." Consider the following examples:

a. In a classic misattribution study, subjects took more shock because they thought a pill caused their symptoms. In a debriefing procedure subjects were asked, "I noticed you took more shock than average. Why do you suppose you did?" Nisbett and Wilson (1977) present one instance of confabulation and claim it as typical. The confabulation begins as follows: "Gee, I don't really know . . ." (p. 237).

b. In a dissonance reduction experiment involving shocks, Zimbardo reports that a typical confabulation would have been, "I guess maybe you turned the shock down" (Nisbett & Wilson 1977, p. 238).

c. Thalia Wheatley, one of the most inventive researchers using hypnotic suggestion (e.g., Wheatley & Haidt 2005), reports that when she has participants perform actions under hypnotic suggestion, she often asks them why they performed the action. Although they do often confabulate, their *initial* response to the question is typically "I don't know" (T. Wheatley, personal communication).

In each of these research paradigms, we find *typical* confabulations delivered with manifestly low confidence, rather than the sense of obviousness and immediacy that is supposed to be characteristic of introspective report.

Carruthers also draws on widely cited cases of confabulation involving split-brain patients. And, although Carruthers claims that split-brain patients confabulate with a sense of obviousness and immediacy, the situation is not so clear. In footage of split-brain patients, we find them showing little confidence when asked to explain behavior issuing from the right hemisphere. For instance, in a typical study with split-brain patient Joe, Joe is shown a saw to his right hemisphere and a hammer to his left. He is then told to draw what he saw with his left hand. Predictably, Joe draws a saw. Gazzaniga points to the drawing and says, "That's nice, what's that?" *Saw*. "What'd you see?" *I saw a hammer*. "What'd you draw that for?" *I dunno* (Hutton & Sameth 1988).

Carefully controlled studies are clearly needed. However, these anecdotes provide prima facie reason to think there are systematic differences in confidence levels between confabulation and apparent introspection, which in turn suggests a difference in underlying mechanism. The fact that confabulations are accompanied by low confidence does not, of course, provide conclusive evidence in favor of introspection. But it does suggest that given the present state of the evidence, the confabulation argument is toothless.

How we know our conscious minds: Introspective access to conscious thoughts

doi:10.1017/S0140525X09000636

Keith Frankish

Department of Philosophy, The Open University, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom.

k.frankish@open.ac.uk

http://www.open.ac.uk/Arts/philos/frankish.htm

Abstract: Carruthers considers and rejects a mixed position according to which we have interpretative access to unconscious thoughts, but introspective access to conscious ones. I argue that this is too hasty. Given a two-level view of the mind, we can, and should, accept the mixed position, and we can do so without positing additional introspective mechanisms beyond those Carruthers already recognizes.

In section 7 of the target article, Carruthers considers the proposal that we have two levels of mentality, conscious and unconscious, corresponding to the two reasoning systems posited by many psychologists, and that we have different forms of access to the attitudes at the two levels – merely interpretative access to those at the unconscious level, but introspective access to those at the conscious level. Prima facie, this mixed position is an attractive one, which does justice both to the evidence for psychological self-interpretation cited by Carruthers and to the everyday intuition that we can introspect our conscious thoughts. Carruthers rejects the option, however. Although conceding that we have introspective access to conscious *thinking*, he denies that we have such access to conscious *judgments* and *decisions*. I argue here that this conclusion is too hasty.

Carruthers' argument turns on the claim that judgments and decisions *terminate* reasoning processes and produce their characteristic effects directly, without further processing. Conscious thinking, on the other hand, involves rehearsing mental imagery, especially inner speech, and this has only an indirect influence on thought and action. The route may be metacognitive: A rehearsed assertion with content p may give rise to an (unconscious) metacognitive belief, to the effect that one believes that p or that one is committed to the truth of p, which, together with suitable desires, will lead one to think and act as if one believes that p. Or the rehearsed assertion may be processed as testimony, leading one to form the first-order belief that p, which will then guide behaviour in the normal way. On either route, Carruthers argues, the conscious event gives rise to the effects of a judgment only through the mediation of further cognitive processing, and so does not count as a judgment itself. Similar considerations apply to decisions, although here Carruthers mentions only the metacognitive route.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

I am sympathetic to Carruthers' account of conscious thinking, and I agree that imagistic rehearsals influence thought and action through the mediation of unconscious cognitive processes. But this is not incompatible with the commonsense view that some conscious events are judgments and decisions. To see this, we need to take seriously the suggestion that the conscious mind is a distinct *level* of mentality. Carruthers has himself developed a version of this view, arguing that the conscious mind (the psychologists' System 2) is not a separate neural structure, but rather, a higher-level "virtual" one, realized in cycles of operation of a more basic unconscious system (System 1), which, among many other tasks, generates and processes the imagery involved in conscious thinking (Carruthers 2006; 2009; for a related version, see Frankish 1998; 2004; 2009). And from this perspective it is natural to regard appropriate utterances in inner speech as genuine judgments and decisions – at least when they achieve their effects via the metacognitive route. For these events will terminate reasoning processes *at the higher level* and *on the relevant topic*. The further processing occurs at the lower level and is devoted to a different topic. When I rehearse the sentence, "Polar bears are endangered" in assertoric mode, this terminates my reasoning about polar bears. The subsequent unconscious reasoning is about how to interpret and respond to this assertion, not about whether the conclusion it expresses is correct. These processes can be thought of as *implementing* the higher-level attitude, and their existence does not compromise the status of the conscious event as a judgment.

It is true that the lower-level processes may sometimes fail to generate the appropriate effects (for example, if the desire to execute one's commitments is overridden by a stronger desire), but this is irrelevant. On every view there are some implementing processes, at least at a neurological level, and these processes may go awry. And if we have a settled habit of interpreting appropriate utterance rehearsals as expressions of belief or commitment, and a settled desire to act consistently or to discharge our commitments, then the right effects will follow most of the time. Similar considerations apply to decisions.

The only peculiarity of the two-level view is that the processes that implement conscious judgments and decisions are cognitive ones. But why should that matter? Compare the way the judgments and decisions of a company are implemented. The edicts emerging from the boardroom require further processing in order to affect the activities of the organization, and this processing involves reasoning on the part of the staff involved. (Again, this will have a metarepresentational character, involving beliefs about what the directors have concluded.) But we still want to say that the judgments and decisions were made in the boardroom, rather than in the cubicles of the junior staff.

What about cases in which a rehearsed utterance generates its effects via the second route, being processed as testimony and generating a first-order belief? Here I think Carruthers is right. If further processing serves to evaluate the conclusion reached rather than simply to implement it, then this does disqualify the conscious event from judgment status. But note that in such cases, the agents themselves will not think of the conscious events as judgments. For if they did, they would naturally come to believe that they believed, or were committed to, the conclusions expressed, and the subsequent processing would follow the metacognitive route. Thus, there is no reason to regard such events as candidates for judgments in the first place. (We might think of them as hypotheses or self-suggestions.) Again, the same goes for decisions.

I conclude that Carruthers' case against a mixed position is not compelling. It is important to stress that the proposed mixed position does not involve positing additional introspective mechanisms. Carruthers allows that we have introspective access to conscious (System 2) thinking; I am simply claiming that some of the introspectable events can be legitimately classified as judgments and decisions. The proposal is merely a reconceptualization of the processes Carruthers describes. But it is a natural one, given a two-level view of the sort Carruthers endorses, and one that accords with intuition. For these reasons it should be preferred. Of course, it would be ad hoc if a two-level view were not independently motivated, but it is (see aforementioned citations).

Non-interpretative metacognition for true beliefs

doi:10.1017/S0140525X09000648

Ori Friedman and Adam R. Petrashek

Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

friedman@uwaterloo.ca

http://www.psychology.uwaterloo.ca/people/faculty/friedman/

arpetras@uwaterloo.ca

Abstract: Mindreading often requires access to beliefs, so the mindreading system should be able to self-attribute beliefs, even without self-interpretation. This proposal is consistent with Carruthers' claim that mindreading and metacognition depend on the same cognitive system and the same information as one another; and it may be more consistent with this claim than is Carruthers' account of metacognition.

Mindreading often requires access to one's own beliefs.[1] Consider the following mental state attributions: Bill *believes* a first-aid kit contains bandages, though the kit actually contains feathers; Louise is an expert in British history, so she *knows* that the Battle of Hastings occurred in 1066; and Sally, age 2, *desires* candy when offered a choice between this and sushi as a snack. These mental state attributions do not depend on the interpretation of others' speech or behavior. Instead, they primarily depend on your beliefs (i.e., first-aid kits normally contain bandages; the Battle of Hastings occurred in 1066; children typically prefer candy over unfamiliar foods) in combination with other principles (e.g., experts in British history know a lot about British history).

The need to access beliefs is not restricted to just a few cases of mindreading. Instead, such access may be the rule in belief attribution: Most beliefs are true, and so one's own beliefs are indicative of what others believe. Because of this, people may have a default tendency to attribute their "true" beliefs to others (Fodor 1992; Leslie & Thaiss 1992; see Leslie et al. [2004] for a review of much evidence favoring an account making this claim). To operate according to this default tendency, the mindreading system requires access to beliefs.

The mindreading system's access to beliefs is problematic for Carruthers' account of metacognition, which denies such access (target article, sect. 2, para. 6).[2] For if the system accesses beliefs when attributing mental states to others, then it should also access them when attributing mental states to the self. For instance, if the mindreading system accesses the belief "the Battle of Hastings occurred in 1066" when attributing it to Louise the historian, then the system should also be able to attribute this belief to the self. The mindreading system's access to beliefs allows people to engage in non-interpretative metacognition.

This proposal does not necessarily imply non-interpretative access to other mental states, such as intentions, desires, and past (currently false) beliefs. Unlike currently held beliefs, these other mental states are typically uninformative about the world and about others' mental states. One's intention to drink coffee says little about the world except perhaps that people sometimes drink coffee; and it says little about other people because relatively few share this intention at any time, meaning that it will seldom be useful to quickly extend this intention to

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

others. So mindreading may not require access to such mental states. If the mindreading system lacks this access, it will also be lacking for metacognition.

Against our proposal, it might be claimed that the mindreading system does not access beliefs, but only inner speech and mental imagery that express beliefs. But this claim requires people to know which fragments of inner speech to use when attributing mental states to others. This claim also contradicts the view that people have a default tendency to attribute true beliefs. And given that inner speech and mental imagery are not required when answering questions about when the Battle of Hastings occurred (sect 2.1, para. 1), it seems doubtful that either is needed when answering when Louise thinks it occurred. Put more baldly, it is difficult to believe that attributing a desire for candy to Sally requires one to express in inner speech the belief "young children typically like candy."

Our proposal is not strongly challenged by evidence that people sometimes confabulate when reporting beliefs. Confabulation is only problematic to the extent that it involves metacognitive errors in which people misreport beliefs. But such errors are difficult to distinguish from accurate reporting of irrational beliefs. When subjects reported that the rightmost of four identical pantyhose was softest (Nisbett & Wilson 1977), they might have been misreporting a belief (i.e., reporting a belief they did not have), but they also might have been faithfully reporting a false belief formed while deciding which item was softest. Also, that people sometimes err in reporting beliefs does not imply that they never have non-interpretative access to their beliefs. Self-interpretation and metacognitive errors may be particularly common for certain sorts of beliefs, and perhaps they are particularly common when people are motivated to report beliefs they do not actually have. In the pantyhose experiment, subjects might have had no belief about which item was softest, but still might have felt compelled to answer. Coming to this answer might open the way for metacognitive errors. But this does not imply that self-interpretation would be needed if subjects were instead asked about something they already believed, such as whether they thought the pantyhose samples were soft at all.

One might also challenge our proposal by conceding that the mindreading system accesses beliefs when making attributions about others, but then denying that it has this access for self-attributions. This defense makes little sense in light of the most detailed account of how beliefs are actually attributed (Leslie et al. 2004). According to this account, the mindreading system operates according to the default assumption that beliefs are true, but sometimes overrides this assumption, as when reasoning about beliefs that are false. This account makes little distinction about whether beliefs are attributed to others or to oneself.

Carruthers' "mindreading is prior" model claims that mindreading and metacognition depend on the same cognitive system and on the same information. Our proposal is consistent with this claim and seems more consistent with it than is Carruthers' account of metacognition. Mindreading requires access to beliefs. Carruthers denies that such access is available in metacognition, which implies that the two processes draw on different information. The account we propose claims that access to beliefs occurs in both mindreading and metacognition, and this implies non-interpretative self-attribution of true belief.

NOTES

1. By access we always mean non-interpretative access. This access might involve a direct link between beliefs and the mindreading system, or it might be indirect and mediated by some other system. We are unsure whether this access conforms to what is normally meant by introspection.

2. Carruthers (2006, especially pp. 181–86) discusses a different version of this problem.

There must be more to development of mindreading and metacognition than passing false belief tasks

doi:10.1017/S0140525X0900065X

Mikolaj Hernik,[a] Pasco Fearon,[b] and Peter Fonagy[c]

[a]*Baby Lab, Anna Freud Centre, London, NW3 5SD, United Kingdom;* [b]*School of Psychology and Clinical Language Sciences, University of Reading, Reading, RG6 6AL, United Kingdom;* [c]*Research Department of Clinical, Educational and Health Psychology, University College London, London WC1E 6BT, United Kingdom.*

mikolaj.hernik@annafreud.org

http://www.annafreudcentre.org/infantlab/mhernik

r.m.p.fearon@reading.ac.uk

http://www.reading.ac.uk/psychology/about/staff/r-m-p-fearon.asp

p.fonagy@ucl.ac.uk

http://www.ucl.ac.uk/psychoanalysis/unit-staff/peter.htm

Abstract: We argue that while it is a valuable contribution, Carruthers' model may be too restrictive to elaborate our understanding of the development of mindreading and metacognition, or to enrich our knowledge of individual differences and psychopathology. To illustrate, we describe pertinent examples where there may be a critical interplay between primitive social-cognitive processes and emerging self-attributions.

Carruthers makes a good case that self-awareness of propositional attitudes is an interpretational process, and does not involve direct introspective access. He also argues that mindreading and metacognition rely on one cognitive mechanism; however, in this case we are less persuaded by the evidence which hinges on Carruthers' reading of well-rehearsed data from autism and schizophrenia. We think that these two predictions have distinct bases and it is at least conceivable that there are two dissociable interpretative meta-representational systems capable of confabulation: one self-directed, one other-directed. Thus, the argument in favour of model 4, over, say, a version of model 1 without a strong commitment to non-interpretative access to self-states, is based purely on parsimony. Our intention is not to defend such a two-system model, but rather to point out that even if one accepts that metacognition involves interpretation, mindreading and metacognition may still be dissociable. Furthermore, Carruthers pays little attention to the differences between input channels associated with first- and third-person mindreading and the surely distinct mechanisms (arguably within the mindreading system) that translate them into attitude-interpretations. As a result, we worry that Carruthers may end up with a rather impoverished model that struggles to do justice to the broader phenotype of first- and third-person mindreading, its development, and the ways in which it may go awry in psychopathology.

Carruthers' reading of developmental evidence is restricted to the standard strategy of comparing children's performance across false-belief tasks. These are inherently conservative tests of mindreading ability, as false-belief-attribution is neither a common nor a particularly reliable function of the mindreading system (Birch & Bloom 2007; Keysar et al. 2003). Clearly, there are earlier and more common abilities central to development of third-person propositional-attitude mindreading – for example, referential understanding of gazes (Brooks & Meltzoff 2002; Senju et al. 2008) or pretense. However Carruthers does not discuss development of the mechanism that is central to his model. He also overlooks evidence that the tendency to engage in pretence has no primacy over the ability to understand pretence in others (Leslie 1987; Onishi et al. 2007).

There are other developmental areas potentially useful to Carruthers' argument. Several socio-constructivist accounts (e.g., Fonagy et al. 2002; 2007) attempt to describe the developmental mechanisms by which early social-cognitive competences, expressed especially in early interactions with the attachment figure (Sharp & Fonagy 2008), give rise to metacognitive awareness. Arguably, the most advanced of these theories is the

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

social-biofeedback model proposed by Gergely and Watson (1996; 1999; Fonagy et al. 2002; Gergely & Unoka 2008). Currently, this model assumes that in repetitive episodes of (mostly) nonverbal communication (Csibra & Gergely 2006) mothers provide marked emotional "mirroring" displays which are highly (but inevitably imperfectly) contingent on the emotional displays of the infant. By doing so, mothers provide specific forms of biofeedback, allowing infants to parse their affective experience, form separate categories of their affective states, and form associations between these categories and their developing knowledge of the causal roles of emotions in other people's behaviour.

It is important to note that socio-constructivist theory is an essential complement to Carruthers' model 4, bridging a potentially fatal gap in his argument. People do *attribute* propositional emotional states to the self, and it seems reasonable to assume that their *actual* emotional states (propositional or not) play a role in generating such attributions. Carruthers' current proposal under-specifies how the mindreading system, which evolved for the purpose of interpreting others' behaviour, comes to be capable of interpreting primary somatic data specific to categories of affective states and of attributing them to the self. Furthermore, according to Carruthers, when the mindreading system does its standard job of third-person mental-state attribution, this sort of data "play little or no role" (target article, sect. 2, para. 8). Presumably, they can contribute, for example, by biasing the outcome of the mindreading processes (like when negative affect leads one to attribute malicious rather than friendly intentions). However, in first-person attributions, their function is quite different. They are the main source of input, providing the mindreading system with cues on the basis of which it can recognize current emotional attitude-states. The social-biofeedback model assumes that the mindreading system is *not readily* capable of doing this job and spells out the mechanism facilitating *development* of this ability. Putting it in terms of Carruthers' model 4: it explains how primary intra- and proprioceptive stimulation gains attentional focus to become globally accessible and how the mindreading system becomes able to win competition for these data.

Research on borderline personality disorder further illuminates the value of the socio-constructivist model (Fonagy & Bateman 2008). The primary deficit in borderline personality disorder (BPD) is often assumed to be a deficit in affect self-regulation (e.g., Linehan 1993; Schmideberg 1947; Siever et al. 2002). We have evidence of structural and functional deficits in brain areas of patients with BPD normally considered central in affect regulation (Putnam & Silk 2005). Accumulating empirical evidence suggests that patients with BPD have characteristic limitations in their self-reflective (metacognitive) capacities (Diamond et al. 2003; Fonagy et al. 1996; Levy et al. 2006) that compromise their ability to represent their own subjective experience (Fonagy & Bateman 2007). There is less evidence for a primary deficit of mindreading (Choi-Kain & Gunderson 2008). Evidence from longitudinal investigations suggests that neglect of a child's emotional responses (the absence of mirroring interactions) may be critical in the aetiology of BPD (Lyons-Ruth et al. 2005), more so even than frank maltreatment (Johnson et al. 2006). We think that the BPD model may become an important source of new data that could illuminate relationships between mindreading and self-awareness and their developmental antecedents. We suggest that children who experience adverse rearing conditions may be at risk of developing compromised second-order representations of self-states because they are not afforded the opportunity to create the necessary mappings between the emerging causal representations of emotional states in others and emerging distinct emotional self-states.

ACKNOWLEDGMENTS

The work of the authors was supported by a Marie Curie Research Training Network grant 35975 (DISCOS). We are grateful for the help and suggestions made by Liz Allison and Tarik Bel-Bahar.

Banishing "I" and "we" from accounts of metacognition

doi:10.1017/S0140525X09000661

Bryce Huebner[a,b] and Daniel C. Dennett[a]

[a]*Center for Cognitive Studies, Tufts University, Medford, MA 02155; and* [b]*Cognitive Evolution Laboratory, Harvard University, Cambridge, MA 02138.*
huebner@wjh.harvard.edu
http://www.wjh.harvard.edu/~huebner
daniel.dennett@tufts.edu
http://ase.tufts.edu/cogstud/incbios/dennettd/dennettd.htm

Abstract: Carruthers offers a promising model for how "we" know the propositional contents of "our" own minds. Unfortunately, in retaining talk of first-person access to mental states, his suggestions assume that a higher-order self is already "in the loop." We invite Carruthers to eliminate the first-person from his model and to develop a more thoroughly third-person model of metacognition.

Human beings habitually, effortlessly, and for the most part unconsciously represent one another *as persons*. Adopting this personal stance facilitates representing others as unified entities with (relatively) stable psychological dispositions and (relatively) coherent strategies for practical deliberation. While the personal stance is not necessary for every social interaction, it plays an important role in intuitive judgments about which entities count as objects of moral concern (Dennett 1978; Robbins & Jack 2006); indeed, recent data suggest that when psychological unity and practical coherence are called into question, this often leads to the removal of an entity from our moral community (Bloom 2005; Haslam 2006).

Human beings also reflexively represent themselves as persons through a process of self-narration operating over System 1 processes. However, in this context the personal stance has deleterious consequences for the scientific study of the mind. Specifically, the personal stance invites the assumption that every (properly functioning) human being is a *person* who has access to *her own* mental states. Admirably, Carruthers goes further than many philosophers in recognizing that the mind is a distributed computational structure; however, things become murky when he turns to the sort of access that we find in the case of metacognition.

At points, Carruthers notes that the "mindreading system has access to perceptual states" (sect. 2, para. 6), and with this in mind he claims that in "virtue of receiving globally broadcast perceptual states as input, the mindreading system should be capable of self-attributing those percepts in an 'encapsulated' way, without requiring any other input" (sect. 2, para. 4). Here, Carruthers offers a model of metacognition that relies exclusively on computations carried out by subpersonal mechanisms. However, Carruthers makes it equally clear that "*I* never have the sort of direct access that my mindreading system has to *my own* visual images and bodily feelings" (sect. 2, para. 8; emphasis added). Moreover, although "*we do* have introspective access to some forms of thinking . . . *we* don't have such access to any propositional *attitudes*" (sect. 7, para. 11; emphasis over "we" added). Finally, his discussion of split-brain patients makes it clear that Carruthers thinks that these data "force us to recognize that *sometimes* people's access to their own judgments and intentions can be interpretative" (sect. 3.1, para. 3, emphasis in original).

Carruthers, thus, relies on two conceptually distinct accounts of cognitive access to metarepresentations. First, he relies on an account of *subpersonal access*, according to which metacognitive representations are accessed by systems dedicated to belief fixation. Beliefs, in turn, are accessed by systems dedicated to the production of linguistic representations; which are accessed by systems dedicated to syntax, vocalization, sub-vocalization, and so on. Second, he relies on an account of *personal access*, according to which *I* have access to the metacognitive representations that allow me to interpret *myself* and form person-level beliefs about *my own* mental states.

The former view that treats the mind as a distributed computational system with no central controller seems to be integral to

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Carruthers' (2009) current thinking about cognitive architecture. However, this insight seems not to have permeated Carruthers' thinking about metacognition. Unless the "I" can be laundered from this otherwise promising account of "self-knowledge," the assumption of personal access threatens to require an irreducible Cartesian *res cogitans* with access to computations carried out at the subpersonal level. With these considerations in mind, we offer what we see as a friendly suggestion: translate all the talk of personal access into subpersonal terms.

Of course, the failure to translate personal access into the idiom of subpersonal computations may be the result of the relatively rough sketch of the subpersonal mechanisms that are responsible for metarepresentation. No doubt, a complete account of metarepresentation would require an appeal to a more intricate set of mechanisms to explain how subpersonal mechanisms can construct "the self" that is represented by the personal stance (Metzinger 2004). As Carruthers notes, the mindreading system must contain a model of *what minds are* and of "the access that agents have to their own mental states" (sect. 3.2, para. 2). He also notes that the mindreading system is likely to treat minds as having direct introspective access to themselves, despite the fact that the mode of access is inherently interpretative (sect. 3.2). However, merely adding these details to the model is insufficient for avoiding the presumption that there must ("also") be *first*-person access to the outputs of metacognition. After all, even with a complete account of the subpersonal systems responsible for the production and comprehension of linguistic utterances, the fixation and updating of beliefs, and the construction and consumption of metarepresentations, it may still seem perfectly natural to ask, "But how do *I know my own mental states*?"

The banality that *I* have access to *my own* thoughts is a consequence of adopting the personal stance. However, at the subpersonal level it is possible to explain how various subsystems access representations without requiring an appeal to a centralized *res cogitans*. The key insight is that a module "dumbly, obsessively converts thoughts into linguistic form and vice versa" (Jackendoff 1996). Schematically, a conceptualized thought triggers the production of a linguistic representation that approximates the content of that thought, yielding a reflexive *blurt*. Such linguistic *blurts* are protospeech acts, issuing subpersonally, not yet from or by the person, and they are either sent to exogenous broadcast systems (where they become the raw material for personal speech acts), or are endogenously broadcast to language comprehension systems which feed directly to the mindreading system. Here, *blurts* are tested to see whether they should be uttered overtly, as the mindreading system accesses the content of the *blurt* and reflexively generates a belief that approximates the content of that *blurt*. Systems dedicated to belief fixation are then recruited, beliefs are updated, the *blurt* is accepted or rejected, and the process repeats. Proto-linguistic *blurts*, thus, dress System 1 outputs in mentalistic clothes, facilitating system-level metacognition.

Carruthers (2009) acknowledges that System 2 thinking is realized in the cyclical activity of reflexive System 1 subroutines. This allows for a model of metacognition that makes no appeal to a pre-existing *I*, a far more plausible account of self-knowledge in the absence of a *res cogitans*.

Unsymbolized thinking, sensory awareness, and mindreading

doi:10.1017/S0140525X09000673

Russell T. Hurlburt

Department of Psychology, University of Nevada, Las Vegas, Las Vegas, NV 89154-5030.

russ@unlv.nevada.edu

http://www.nevada.edu/~russ

Abstract: Carruthers views unsymbolized thinking as "purely propositional" and, therefore, as a potential threat to his mindreading-is-prior position. I argue that unsymbolized thinking may involve (non-symbolic) sensory aspects; it is therefore not purely propositional, and therefore poses no threat to mindreading-is-prior. Furthermore, Descriptive Experience Sampling lends empirical support to the view that access to our own propositional attitudes is interpretative, not introspective.

Section 8 of Carruthers' target article considers my Descriptive Experience Sampling (DES) work, particularly its finding of unsymbolized thinking (Hurlburt 1990; 1993; 1997; Hurlburt & Akhter 2008; Hurlburt & Heavey 2006). Carruthers implies that I characterize unsymbolized thinking as being purely propositional: "many subjects also report the presence of 'purely propositional,' unsymbolized thoughts at the moment of the beep" (sect. 8, para. 2). As a result, he supposes that my claim that unsymbolized thoughts are introspected (Hurlburt 1990; 1993) might present a difficulty for his mindreading-is-prior view, which holds that purely propositional events are not introspected but are, instead, interpreted.

Against this supposition, Carruthers argues that the introspection of unsymbolized thinking is an illusion; what is mistaken for introspection is a swift but unconscious interpretation of external events (Carruthers 1996b) and/or internal events such as images (present target article). As a result, he concludes in the target article that DES is neutral regarding Carruthers' mindreading view: "although there is no *support* to be derived for a 'mindreading is prior' account from the introspection-sampling data, neither is there, as yet, any evidence to count against it" (sect. 8, para. 5, emphasis in original).

I think Hurlburt and Akhter (2008) successfully rebutted Carruthers (1996b), and the target article does not change my mind. But I agree that unsymbolized thinking does not threaten Carruthers' mindreading-is-prior position, not because unsymbolized thinking is an unconscious interpretation but because it is not "purely propositional." Unsymbolized thinking is a directly apprehendable experience that may well have some kind of (probably subtle) sensory presentation, is therefore not purely propositional, and for that reason is not at odds with the mindreading-is-prior view.

In seeking to discover why Carruthers might hold, mistakenly, that I believe that unsymbolized thinking is "purely propositional," I reviewed what I have written on unsymbolized thinking and discovered this sentence:

> Unsymbolized Thinking is the experience of an inner process which is clearly a thought and which has a clear meaning, but which seems to take place without symbols of any kind, that is, *without* words, images, *bodily sensations*, etc. (Hurlburt 1993, p. 5; emphasis added)

"Without ... bodily sensations" might be understood to mean "purely propositional," but that is not at all what I intended. I should have written "without ... bodily *sensory awareness*" instead of "without ... bodily *sensations*."

"Sensory awareness" is a term of art in DES: "A sensory awareness is a sensory experience (itch, visual taking-in, hotness, pressure, hearing) that is in itself a primary theme or focus for the subject" (Hurlburt & Heavey 2006, p. 223). That is, sensory awareness is not merely a bodily or external sensation, but is a sensation that is itself a main thematic focus of experience. Thus, for example, Jack picks up a can of Coke, and, while preparing to drink, particularly notices the cold, slippery moistness against his fingertips. Jill picks up a can of Coke, and, while preparing to drink, says to herself in inner speech, "Carruthers is right!" Both Jack and Jill are having bodily sensations of the coldness, the moistness, and the slipperiness of the can (neither drops it). Jack's central focus is on the cold, slippery moistness; therefore, he is experiencing a sensory awareness as DES defines it. Jill's central focus is on her inner speech, not on the can; therefore she is *not* experiencing a sensory awareness as defined by DES (see Hurlburt & Heavey, in preparation).

Thus, unsymbolized thinking, as I and my DES colleagues describe the phenomenon, is an experience that is directly apprehended at the moment of the DES beep but which does not

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

involve the direct apprehension of verbal, imaginal, or other symbols and does not involve sensory awareness as DES defines that term. The apprehension of an unsymbolized thought may involve the apprehension of some sensory bits, so long as those sensory bits are not organized into a coherent, central, thematized sensory awareness. Thus, I believe that unsymbolized thinking is a perceptual event, just as are inner speech, visual imagery, and feelings; it is therefore not purely propositional and is therefore not a threat to the mindreading-is-prior view.

Access to propositional attitudes is interpretative. Far from being neutral, DES lends empirical support to the main thrust of Carruthers' analysis that propositional attitudes are interpreted, not observed. The DES procedure trains subjects carefully, repeatedly, and iteratively (Hurlburt & Akhter 2006; Hurlburt & Heavey 2006; Hurlburt & Schwitzgebel 2007) to distinguish between directly observed (Carruthers' "perceptual") events and all else; that training typically requires several days. DES tries, moment by moment, to cleave to the directly observed and to bracket all that is inferred, supposed, presupposed. There is no a priori assumption about what is or is not directly observable. Attitudes are not singled out; if an attitude is directly observed at the moment of some beep, then that attitude is the proper target of DES. If not, then it isn't.

As a result of 30 years of carefully questioning subjects about their momentary experiences, my sense is that trained DES subjects who wear a beeper and inspect what is directly before the footlights of consciousness at the moment of the beeps almost never directly apprehend an attitude. Inadequately trained subjects, particularly on their first sampling day, occasionally *report* that they are experiencing some attitude. But when those reports are scrutinized in the usual DES way, querying carefully about any perceptual aspects, those subjects retreat from the attitude-was-directly-observed position, apparently coming to recognize that their attitude had been merely "background" or "context." That seems entirely consonant with the view that these subjects had initially inferred their own attitudes in the same way they infer the attitudes of others. (I note that subjects do not similarly retreat from their initial reports about unsymbolized thinking; they continue to maintain that the unsymbolized thought had been directly observed.)

What monkeys can tell us about metacognition and mindreading

doi:10.1017/S0140525X09000685

Nate Kornell,[a] Bennett L. Schwartz,[b] and Lisa K. Son[c]

[a]*Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095-1563;* [b]*Department of Psychology, Florida International University, Miami, FL 33199;* [c]*Department of Psychology, Barnard College, New York, NY 10027.*

nkornell@ucla.edu **http://nkornell.bol.ucla.edu/**
bennett.schwartz@fiu.edu **www.fiu.edu/~schwartb**
lson@barnard.edu **http://lisason.synthasite.com/index.php**

Abstract: Thinkers in related fields such as philosophy, psychology, and education define metacognition in a variety of different ways. Based on an emerging standard definition in psychology, we present evidence for metacognition in animals, and argue that mindreading and metacognition are largely orthogonal.

The target article proposes that "mindreading is prior to metacognition," meaning that just as we know the minds of others by observing what they do, we know our own minds by observing what we do. According to this view, metacognition – that is, cognition about one's own cognition – requires mindreading abilities. Rhesus monkeys (*Macaca mulatta*) do not appear to possess mindreading abilities (Anderson et al. 1996; but see Santos et al. 2006). Here we present evidence, however, that rhesus monkeys are metacognitive. We offer a different definition of mindreading than that used by Carruthers, and we contend that the mechanisms of mindreading and metacognition are largely orthogonal.

The target article reports in detail on only a few seminal studies of metacognition in animals (see Smith et al. 2003; Smith & Washburn 2005). We begin by elaborating on subsequent studies that provide evidence of animal metacognition (reviewed by Kornell, in press). For example, Hampton (2001) tested monkeys in a modified delayed match-to-sample task: On each trial, a sample picture was presented on a touch-sensitive computer monitor, and then, after a delay, the same sample picture was presented among three distractors, and the subject had to touch the sample. On some trials, after viewing the sample, the monkey could choose to skip the test and receive a small reward. If the monkey instead chose to take the test, he could earn a large reward, or, if his response was incorrect, forfeit reward completely. Memory accuracy was better on self-selected test trials than on mandatory test trials. It appears that the monkeys chose to take the test when they knew that they knew the answer, in the same way that a student raises her hand in class when she knows that she knows the answer (see Suda-King 2008, for similar results in orangutans).

In another study, two male rhesus monkeys were asked, essentially, to bet on their memories (Kornell et al. 2007). A given monkey was shown six pictures sequentially for "study," followed by a display of nine pictures presented simultaneously, one of which had been "studied." The monkey's task was to select the studied picture. After he responded, two "risk" icons were presented, which allowed the monkey to bet his tokens (which could be exchanged for food). A high-risk bet resulted in the gain of three tokens if the monkey had responded correctly, but a loss of three tokens otherwise. Choosing low-risk resulted in a sure gain of one token. The monkeys made accurate confidence judgments: They bet more after correct responses than after incorrect responses. This finding was especially impressive because the monkeys were originally trained on tasks that involved neither pictures nor remembering (e.g., select the longest line); following that training, they were able to respond metacognitively beginning on the first day of the picture-memory task. The monkeys appear to have learned a general metacognitive response, not one that was task-specific.

In addition to being able to make judgments about their memories, monkeys have demonstrated that they can choose behaviors, based on metacognition, that advance their own knowledge – that is, they have demonstrated metacognitive control (see Nelson & Narens 1990). To investigate this ability, we allowed two monkeys to request information when they were uncertain, just as a person might ask for a hint when answering a difficult question (Kornell et al. 2007). The monkeys could request a "hint," that is, a blinking border that surrounded the correct response, on some trials in a list-learning experiment. As the monkeys' response accuracy on no-hint trials improved steadily, their rate of hint requests showed a corresponding decline. By requesting hints when they were unsure, the monkeys went beyond making an uncertain response; they took steps to rectify their ignorance.

Based on the studies described above, we conclude that monkeys have metacognitive abilities – that is, they can monitor the strength of their own internal memory representations. According to the target article, these findings fall short of metacognition, however. Carruthers writes, "It is only if a human reports that she acted as she did, not just because she *was* uncertain, but because she was *aware of being* uncertain, that there will be any conflict [with the metacognition is prior account]" (sect. 5.2, para. 3). We do not agree that metacognition requires awareness; we have previously argued that the metacognitive abilities that animals possess are not necessarily conscious (Kornell, in press; Son & Kornell 2005; also see Reder 1996).

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

For example, a monkey might make a high-risk bet without being aware that it is monitoring its internal memory trace.

We are not arguing that mindreading cannot subsume metacognitive functions. Indeed, we can learn much about ourselves by observing our own behavior: for example, after playing a round of golf, we decide we are not quite ready for the pro tour. Moreover, numerous experiments have shown that metacognition is largely based on unconscious inferential processes, not direct examination of memories; for example, we infer that we know something well based on the fluency (i.e., ease and speed) with which it comes to mind (Schwartz et al. 1997).

Given the way we, and many other cognitive psychologists, define metacognition, we assert that it is likely that metacognition and mindreading are separate processes. The argument that one should only see metacognition in species that can mindread is, to the best available evidence, false. For example, some have suggested that dogs, which have shown no metacognitive abilities but show high levels of social cognition, may have rudimentary mindreading abilities (Horowitz, in press; Tomasello et al. 1999). Conversely, we offer rhesus monkeys as a case study in a metacognitively competent animal that fares poorly at mindreading. In the tasks we describe, metacognitive processing can lead to positive outcomes that are evolutionarily adaptive. Indeed, metacognitive monitoring seems to have its own rewards.

Metacognition without introspection

doi:10.1017/S0140525X09000697

Peter Langland-Hassan

Department of Philosophy, The Graduate Center of the City University of New York, New York, NY 10016.

PLangland-Hassan@gc.cuny.edu

https://wfs.gc.cuny.edu/PLangland-Hassan

Abstract: While Carruthers denies that humans have introspective access to cognitive attitudes such as belief, he allows introspective access to perceptual and quasi-perceptual mental states. Yet, despite his own reservations, the basic architecture he describes for third-person mindreading can accommodate first-person mindreading without need to posit a distinct "introspective" mode of access to *any* of one's own mental states.

Carruthers argues that passivity symptoms (e.g., thought insertion) in schizophrenia result not from a special metacognitive deficit, but from "faulty data being presented to the mindreading system" (sect. 9, para. 2). Although I endorse Carruthers' Frith-inspired (Frith et al. 2000a; 2000b) appeal to efference-copy deficits in the explanation of passivity symptoms, his claim that the mindreading faculty itself is undamaged raises questions. First, any attribution of one's own thoughts to another is equally a mistake in first- and third-person mindreading (false positives count as errors just as much as false negatives do). Carruthers should therefore hold that mindreading – first- *and* third-person – is deficient in these forms of schizophrenia; this still allows him to deny any dissociation between mindreading and metacognitive abilities, in line with what his theory predicts. It also avoids his having to make the hard-to-test claim that it is intermittently faulty data and not an intermittently faulty mechanism that is to blame for passivity symptoms.

Second, Carruthers holds that humans have introspective access to some mental states (e.g., perceptual states, imagery, and inner speech), but not to cognitive attitudes such as belief. But if information is extracted from globally broadcast perceptual states in third-person mindreading without introspection occurring, why think that the extraction of information from inner speech and visual imagery during first-person mindreading involves an introspective process different "in kind" from the way we form beliefs about the mental states of others? If, as Carruthers argues, passivity symptoms result from faulty data being input to the mindreading system (data that should have been interpreted as internally generated is interpreted as externally generated), then it seems the very determination of whether an input is self or other-generated – and thus whether one is seeing or visualizing, hearing or sub-vocalizing – requires an inferential or interpretative step (Langland-Hassan 2008).

Carruthers would likely respond that this inner-or-outer inferential step involves nothing more than the "trivial" form of inference that occurs in any layered representational scheme, where representations at one level can, in a "supervisory" role, intervene on those at another. However, many instances of third-person mindreading are equally fast and automatic, and they are implicit in the very cases of metacognition that, on Carruthers' theory, would be achieved through the "encapsulated" process of introspection. Consider a visual representation had by someone who looks up and sees another person staring at him. Suppose this visual perceptual state is accessed by the mindreading system, which issues in the introspective judgment: "I see a man seeing me." This judgment contains within it a judgment that another person is having a visual experience of a certain kind (cf. Jeannerod & Pacherie's [2004] "naked intentions"). So, unless the mindreading faculty in its introspective mode lacks the concepts needed for this judgment (unlikely, since it must have the concepts of self and of sight in order to issue *any* introspective judgments about visual experience), third-person mindreading can occur through the encapsulated "introspective" process that Carruthers describes. Yes, some cases of third-person mindreading require much more sophisticated feats of interpretation, but so too do many cases of first-person mindreading, as revealed by the confabulation data Carruthers discusses (Gazzaniga 1995).

Thus, even if it is possible to draw a line between mindreading that is informationally encapsulated and that which is not, it will not cut cleanly across cases of first- and third-person mindreading. Nor is the existence of such domain-specific mechanisms supported by recent neuroimaging studies (Decety & Lamm 2007). What we have instead are inferences, concerning both first- and third-person mental states, that require greater or lesser degrees of supporting information; none of this implies a special *mode* of access to facts about one's own mental states. This is obscured by the tendency of researchers to compare easy cases of metacognition (e.g., inferring one's intentions from one's own inner speech) with difficult cases of third-person mindreading (e.g., inferring what someone thinks based solely on their posture and facial expression) – for it creates the impression that first-person mindreading occurs through some more "direct" process. But if we instead compare the third-person mindreading that occurs when we judge that a person believes what we hear her saying, to the first-person mindreading that draws on "listening" to one's own inner speech, there is less intuitive pressure to posit a difference in the kind of inference. Of course, if there were genuine dissociations revealed between third- and first-person mindreading abilities, as Nichols and Stich (2003) and Goldman (2006) claim, then we would have reason to posit differences in the kinds of mechanisms and inferences involved in each; but Carruthers is at pains to deny any such dissociations, and his alternative explanations are plausible enough.

The issue can be reframed in terms of the larger evidence base we have for first-person rather than third-person mindreading. Carruthers notes that the resources available to first-person mindreading are different because, "unless subjects choose to tell me, I never have access to what they are imagining or feeling" (sect. 2, para. 8). This is potentially misleading; the situation is rather that the single mindreading system, as he describes it, *only ever* has access to globally broadcast perceptual and quasi-perceptual representations (and memory), and, with this single source of information, must accomplish both its first- and third-person mindreading tasks – one of which is to determine whether the signal counts as a case of imagining or perceiving in the first place.

The fact that we have so much more "evidence" for first-person mindreading than third-person may still tempt some to posit

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

a special form of access. Yet, if humans always audibly narrated their inner speech and expressed the contents of their visual imagery, the evidence bases for first- and third-person mindreading would be comparable. So it may be a contingent fact about how humans behave that accounts for the difference in evidence bases, not a special mode of access.

I therefore urge Carruthers to adopt a more thoroughgoing anti-introspectionism. Not only can first-person mindreading be explained without appeal to the introspection of propositional attitudes, it can be explained without granting a distinct introspective form of access to *any* of one's own mental states.

Carruthers' marvelous magical mindreading machine

doi:10.1017/S0140525X09000703

Charlie Lewis[a] and Jeremy I. M. Carpendale[b]

[a]*Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster, LA1 4YF, United Kingdom;* [b]*Department of Psychology, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.*

c.lewis@lancsater.ac.uk

http://www.psych.lancs.ac.uk/people/CharlieLewis.html

jcarpend@sfu.ca

http://www.psyc.sfu.ca/people/faculty.php?topic=finf&id=67

Abstract: Carruthers presents an interesting analysis of confabulation and a clear attack on introspection. Yet his theory-based alternative is a mechanistic view of "mindreading" which neglects the fact that social understanding occurs within a network of social relationships. In particular, the role of language in his model is too simple.

In his analysis of four identifiable accounts of the relationship between "mindreading" and "metacognition," Carruthers makes a bold attempt to explain a major inconsistency in three of these and assert the primacy of "mindreading" as a precursor to self-understanding. He does this mainly by highlighting a common key feature in the literature in neuropsychology, reasoning, and psychopathology: why adults with sophisticated skills in accounting for their own and others' thoughts still seem to show flaws in their abilities to demonstrate and reconcile these skills. However, we make three connected points to argue that his account falls short of a clear answer to his main concerns about the nature of the two processes and how they relate.

First, Carruthers treats the terms "metacognition" and "mindreading" as unproblematic, distinct, and clearly grounded within an overarching, internalist cognitive architecture. Yet, the article is unclear about how this architecture fits together into a working psychological system. There is more than a hint of a homunculus in Carruthers' descriptions of how we access and use these skills (Kenny 1971/1991). Let's look at the role of the role of language in mindreading as an example.

Early in the target article Carruthers dismisses the importance of language in a note because it seems "plausible" to him "that basic capacities for both mindreading and metacognition are independent of our capacity for natural language" (target article, Note 2). Yet in his analysis of reasoning (sect. 8), he acknowledges its importance in certain circumstances, but does not convincingly show how this contradiction can be reconciled. Then, in the bulk of the article, speech only serves as input for Carruthers' "magical mindreading machine," revealing a view that thinking is computation and meaning is mechanistic, entailing a code model of language. All of these assumptions have been repeatedly critiqued (e.g., Bickhard 2001; Goldberg 1991; Heil 1981; Proudfoot 1997; Tomasello 2003; Turnbull 2003; Wittgenstein 1968). For example, problems have been pointed out with how the symbols involved in computation could acquire meaning, and whether understanding can adequately be conceived of as a process of computation (Proudfoot 1997). These criticisms have been applied to the theoretical approach that Carruthers advances.

Carruthers claims to have considered four different theories. However, since he starts with such a restricted set of theories to consider, knowing that he declares his version of theory-theory as the winner in this contest tells us nothing about how the many other approaches would fare. The *one* place where we *do* agree with Carruthers is that assuming introspection as a *source* of social understanding is problematic. We find it odd, however, that he ignores Wittgenstein's far stronger arguments against introspection. Alternative accounts based upon a more sophisticated grasp of language deal more easily with the confabulation literature that is so well described in the target article (sect. 6), but not fully explained.

Indeed, second, the theory-theory's response to these criticisms is simply to ignore it (Carpendale & Lewis 2004; Leudar & Costall, in press). Carruthers' perspective is symptomatic of an insularity that has been commented upon repeatedly in the wider literature, but which barely gets a mention in the "mainstream." The idea of the 1980s that we, or something inside ourselves, can manipulate our own social understanding was criticised by Bruner (1986; 1990), who rejected a cognitive system that is bereft of serious contact with the outside world. Many commentators have repeated this claim. Carruthers buys into forms of attributing mental states to ourselves and others which assume the theory metaphor so loosely as to be bereft of meaning (see Campbell & Bickhard 1993). The fact that this movement has been as immune to debate in the outside world as the cognitive system that they depict has led its critics (e.g., Leudar & Costall, in press, Ch. 1) to use the term *ToMism* to refer to the assumption that the individual is trapped behind a mindreading mechanism that filters their interpretation and observations of social interactions.

Third, in his neglect of the growing literature which attempts to explore the interactional basis of social understanding, Carruthers' account is both parasitic upon the evidence from developmental psychology and simultaneously dismissive of its analysis of the nature of the two processes. We suspect that without a clear analysis of development of "mindreading" and/or "metacognition," Carruthers' definition of these two terms remains grossly underspecified. There is a large body of recent work that attempts to explore the dynamics of interaction in early (Reddy 2008) and later infancy (Moll & Tomasello 2007) and in childhood (Carpendale & Lewis 2004), as well as with reference to special populations (Hobson 2002), which describes how social understanding, as opposed to some miraculous yet unspecified mindreading system, gradually emerges in social interaction. In these approaches there is no Cartesian split between the individual in the spectator's role observing others' physical movements and having to attribute mental states. Without a reference to this expanding and exciting literature, and its data and critique of the solipsism of theory-theory, Carruthers' argument is a better description of atypical "mindreading" like autism, not an account of how we understand ourselves and others.

What neuroimaging and perceptions of self-other similarity can tell us about the mechanism underlying mentalizing

doi:10.1017/S0140525X09000715

Michael V. Lombardo, Bhismadev Chakrabarti, and Simon Baron-Cohen

Autism Research Centre, University of Cambridge, Cambridge CB2 8AH, United Kingdom.

ml437@cam.ac.uk bhisma@cantab.net sb205@cam.ac.uk

http://www.autismresearchcentre.com

Abstract: Carruthers' "mindreading is prior" model postulates one unitary mindreading mechanism working identically for self and other.

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

While we agree about shared mindreading mechanisms, there is also evidence from neuroimaging and mentalizing about dissimilar others that suggest factors that differentially affect self-versus-other mentalizing. Such dissociations suggest greater complexity than the mindreading is prior model allows.

The "mindreading is prior" model proposed by Carruthers postulates that one mechanism (mindreading) is all that is needed to understand the mental states of oneself and others. Although we agree that shared mechanisms can implement the computations underlying both self- and other-mentalizing, we question whether all types of mentalizing use this one mechanism in an identical fashion. We present evidence from functional neuroimaging and research suggesting that mentalizing differs when the target individuals are of differing similarity to oneself.

Because the "mindreading is prior" model asserts that one mechanism is used indifferently for self- and other-mentalizing, this generates three predictions that all need to be satisfied in the brain. First, there should be an identical neural implementation of mentalizing for self and other. Thus, the *same* brain regions that code for self-mentalizing computations should also be recruited for other-mentalizing. Second, such identical neural implementation should occur to a similar degree for both self and other. Thus, activation in shared regions should be *equal* in self- and other-mentalizing. Finally, there should be *no other* areas in the brain that are recruited specifically for just self- or just other-mentalizing.

In our own recent work (Lombardo et al., submitted) we find evidence that while predictions 1 and 3 are satisfied, prediction 2 is not. We find that although the ventromedial prefrontal cortex (vMPFC), posterior cingulate/precuneus (PCC), and right temporo-parietal junction (RTPJ) are all recruited for mentalizing about both self and other, they do so in a target-biased fashion. The vMPFC is recruited more for self-mentalizing than other-mentalizing, whereas PCC and RTPJ are recruited more for other- than self-mentalizing. Thus, while it is the case the identical neural mechanisms are implementing self- and other-mentalizing computations, they do not do so in an equal fashion. Some of these shared neural representations are tailored more for handling self-mentalizing, whereas other shared representations are tuned in to handle other-mentalizing. The "mindreading is prior" model is silent about how one unitary mindreading mechanism can be tuned preferentially for mentalizing about self more than other, or vice versa.

In addition to these neuroimaging results, there is behavioral evidence in individuals with autism spectrum conditions (ASC) that the magnitude of impairment in self- and other-mentalizing is unequal (Lombardo et al. 2007). In the domain of emotion understanding, on measures of other-mentalizing involving reading complex emotions (Reading the Mind in the Eyes Test; Baron-Cohen et al. 2001), adults with ASC show less impairment (Cohen's $d = 0.61$) than on measures assessing understanding one's own emotions (Toronto Alexithymia Scale; Cohen's $d = 1.40$). Thus, although impairment may exist in both self- and other-mentalizing in ASC, the impairments are *unequal* in magnitude. This evidence presents a paradox: How can one unitary mechanism working identically for self and other affect self and other in a differential manner?

The "mindreading is prior" model is also silent about what happens when an individual mentalizes about an individual of varying similarity to oneself. Carruthers treats the "other" as completely distinct from oneself, and he bases his theory on such a "monolithic other." However, social psychological research suggests that the mechanisms for mentalizing about others functions differently depending the degree to which an individual perceives the other person to be similar or dissimilar to oneself. When another is perceived to be dissimilar to oneself, mentalizing responses follow a pattern similar to stereotyping; we expect dissimilar individuals to share mental states of individuals from the salient group that they belong to. However, when another individual is similar to oneself, social inference proceeds in a congruent manner to what we ourselves think or feel (Ames 2004).

A similar distinction can be seen within neuroimaging research on this topic. Jason Mitchell and colleagues (Jenkins et al. 2008; Mitchell et al. 2006) present elegant work showing that similar neural responses occur in vMPFC during self-mentalizing and mentalizing about similar others, but not during mentalizing about dissimilar others. Mentalizing about dissimilar others involves processing in the dorsomedial prefrontal cortex (dMPFC) rather than the vMPFC. Furthermore, when stereotyping is applied in the context of mentalizing about others, areas involved in semantic retrieval and selection such as the ventrolateral prefrontal cortex (VLPFC) are recruited, but not the vMPFC or dMPFC (Mitchell et al., in press).

Thus, behavioral and neural distinctions can be made about the mechanism underlying mentalizing about others of differing similarity to oneself. In the case of similar others, the regions involved in both self- and other-mentalizing may overlap considerably. However, different regions appear to be at work when making inferences about the mental states of dissimilar others. How does the "mindreading is prior" model account for such differences in the mindreading mechanism, solely based on the perceived similarity of others to oneself?

Finally, in our own work, we find that even in ASC the mechanism underlying mentalizing about self and similar others may be different to mentalizing about dissimilar others (Lombardo et al. 2007). We asked participants to make judgments about personality traits (e.g., *honest, caring, anxious*) in relation to themselves, a similar close other, and a dissimilar non-close other. In a control task, they were asked to simply count the syllables in personality trait words. In a surprise recognition memory test, we found that individuals with ASC had no impairment in memory for words encoded by syllable counting or in relation to a dissimilar other. However, when looking at memory for self-encoded or similar other-encoded traits, we found substantial impairment. Individuals with ASC had an impairment in processing linked to thinking about themselves and similar others, but no deficit in regards to a dissimilar other. Such dissociations for similar and dissimilar others imply the mindreading mechanism may conceal a greater complexity than is suggested by Carruthers' model.

Feigning introspective blindness for thought

doi:10.1017/S0140525X09000727

Robert W. Lurz

Department of Philosophy, Brooklyn College, City University of New York, Brooklyn, NY 11218.

rlurz@brooklyn.cuny.edu

http://depthome.brooklyn.cuny.edu/philo/Lurz.htm

Abstract: I argue that the very reasons Carruthers gives for why the "mindreading is prior" account should allow introspective access to perceptual/quasi-perceptual states, can be given for thought, as well. I also argue that we have good subjectively accessible grounds for the intuition in introspective thoughts, notwithstanding Carruthers' argument to the contrary and his attempt to explain the intuition away.

I. Carruthers argues that a consequence of the "mindreading is prior" account is that the mindreading faculty should have introspective access to perceptual and quasi-perceptual states. Two reasons are given. First, because the mindreading faculty must have access to perceptual inputs about the actions of others (targets), it "should be capable of self-attributing those percepts in an 'encapsulated' way, without requiring any other input"

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

(target article, sect. 2, para. 4). But arguably the mindreading faculty would not have evolved unless it were able to *predict* the behaviors of others; and for this, the faculty must have access to non-perceptual beliefs about past and general facts about particular targets. If a mindreading animal, for example, observes a target orient its eyes toward food and thereby attributes *sees-food*, it is unlikely that it will be able to predict what the target will do unless it has access to relevant past and general facts about the target that are not perceptually available in the scene or hardwired into the faculty itself (e.g., that the target is a member of its own group/family, has recently eaten, has shared food in the past, etc.). By parity of reasoning, since the mindreading faculty must have access to non-perceptual beliefs, the faculty should be capable of self-attributing such thoughts in an "encapsulated" way without requiring any further input.

Carruthers argues that introspection of quasi-perceptual states is also likely on the "mindreading is prior" account because, being perceptual, they are (when attended to) "globally broadcast" to all concept-forming systems, including the mindreading faculty. But arguably attended-to thoughts are "globally broadcast" as well. In fact, Baars (1997) argues that attended-to thoughts, just as much as attended-to percepts, "create a vast access to perhaps all parts of the nervous system" (p. 59). For example, if one were to observe a subject looking at a red apple, one's mindreading faculty would likely infer that the subject sees the color of the apple; however, this default inference would have been prevented if, prior to observing the subject, the thought should have occurred to one that this subject had recently informed one that she was color-blind. It is quite plausible, therefore, that attended-to thoughts are also broadcasted to the mindreading factually, and by parity of reasoning, the faculty should be capable of self-attributing such thoughts in an "encapsulated" way without requiring any further input.

II. The standard argument in support of introspective access to thoughts runs as follows (see Goldman 2006, p. 230, for an example):

1. Sometimes we know what we think, and yet what we think is quite unrelated (for interpretative purposes) to any of the contents to which we have subjective access (such as the contents of perception, proprioception, episodic memory, or the contents of introspection regarding perceptual and quasi-perceptual states).
2. Therefore, it is unlikely in such cases that we know what we think as a result of an interpretation from the subjectively accessible contents rather than as a result of an introspective access to the thought itself.

This is an inductive argument to the best explanation. Hence, the more interpretatively irrelevant the subjectively accessible contents are to the self-ascribed thought, the more likely the process is introspection. Furthermore, because the processes involved in interpretation and introspection are unconscious, it is to be expected that the greater the degree of interpretative relevance the subjectively accessible contents bear to the self-ascribed thought, the greater the chances are that it will appear to the subject as if he knows what he thinks on the basis of introspection when it is actually the result of interpretation. I suspect that this is what is happening in the split-brain case that Carruthers describes. The subject has access to the contents of his perceptual and proprioceptive states that represent him as walking out of the testing van, and he probably has access to the contents of his memory which represent the location of the van to the house, and that the house has a Coke in it, and so on. All of this would likely allow him to interpret (at an unconscious level) himself as going to get a Coke from the house. Is it possible for it to *always* appear as if our knowledge of our own thoughts is introspective when it is interpretative? Yes, but it is unlikely the more interpretatively unrelated these contents are to the self-ascribed thought.

The preceding argument assumes that the belief in introspective thoughts is the result of an inductive argument and is not simply built into our pre-theoretic concept of the mind's epistemic access to itself. Carruthers disagrees. He argues that on the "mindreading is prior" account, it is to be expected that the mindreading system should represent this epistemic access as introspection and not interpretation, since doing so would "greatly simplify" the system's computational operations without any significant loss in reliability. However, if the mindreading system's concept of introspection is Carruthers' broad and negative definition (roughly, a reliable method for forming metacognitive judgments that is *not* interpretative and is different in kind from our access to other minds), then in order for the mindreading system to apply its concept of introspection to a subject, it will need to assume or have reason to believe that the subject's access to its mind is *not* interpretative. This would seem to undermine Carruthers' claim of there being a greater degree of simplification in the operations of a mindreading system that represented the mind's access to itself as introspection compared to one that represented it as interpretation only, since the former would require the possession and application of the concepts of introspection *and* interpretation, whereas the latter would require the possession and application of the concept of interpretation only. It is more plausible to suppose that the mindreading system's model of the mind's access to itself would simply be that of a reliable judgment-forming process that is different in kind from that used to form judgments about other minds. But such a model of the mind's access to itself would be neutral regarding whether it is introspection (as Carruthers defines it) or interpretation – at least, with respect to the kind of interpretivism that Carruthers defends in the "mindreading is prior" account, which also holds that the mind's interpretative access to its own thoughts is reliable and different in kind from its access to the thoughts of others.

Getting to know yourself . . . and others

doi:10.1017/S0140525X09000739

Candice M. Mills[a] and Judith H. Danovitch[b]

[a]*School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX 75083-0688;* [b]*Department of Psychology, Michigan State University, East Lansing, MI 48824-1116.*

candice.mills@utdallas.edu

http://www.utdallas.edu/research/thinklab

jhd@msu.edu

http://psychology.msu.edu/people/faculty/danovitch.htm

Abstract: Carruthers rejects developmental evidence based primarily on an argument regarding one skill in particular: understanding false beliefs. We propose that this rejection is premature; and that identifying and examining the development of other subcomponent skills important for metacognition and mindreading, such as the ability to assess levels of knowledge, will in fact be useful in resolving this debate.

Although we find his proposal thought-provoking, we disagree with Carruthers' conclusion that "developmental evidence is of no use" (target article, sect. 4, para. 3) in determining the relationship between mindreading and metacognition, because it is based primarily on an incorrect assumption that false belief performance is the only relevant measure of mindreading and metacognition in children. False beliefs tasks are just one way of measuring mindreading, and not necessarily a good way (Bloom & German 2000). Indeed, most developmental psychologists believe that mindreading and metacognition involve a number of subcomponent skills. Some of these skills, such as the ability to understand that others have intentions, may be mastered by infants and toddlers (e.g., Meltzoff 1995), whereas

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

others, such as the ability to appreciate false beliefs, are mastered in the early preschool years (Wellman et al. 2001). Other skills, such as the ability to understand that a person can have conflicting desires (e.g., Choe et al. 2005) or the ability to accept that two people can interpret the same information in different ways (e.g., Carpendale & Chandler 1996), are not grasped until middle childhood. Indeed, many of the component skills that enable mindreading and metacognition are unaccounted for in Carruthers' account of why developmental evidence is not useful. Thus, by examining self/other comparisons in the development of these skills, developmental data can help us better understand potential asymmetries in the emergence of mindreading and metacognition.

Carruthers identifies one possible area for further examination. In his introduction, he points out that mindreading involves not only intentions and beliefs (which are measured in the false belief task), but also components such as knowledge. However, his review of the developmental literature does not take into account the growing body of research examining young children's assessments of what other people know versus what they themselves know.

Determining what another person is likely to know can be considered mindreading because it requires an understanding of the contents and limitations of that person's thoughts. In order to make a judgment about what another person knows, children cannot always rely on associations or broad generalizations. Rather, the evidence suggests that young children have a sophisticated understanding of what other people know or do not know. For instance, they understand that one person can be knowledgeable about some things but not knowledgeable about other things (e.g., Harris 2007; VanderBorght & Jaswal, in press). Even 3-year-olds, who might be considered poor at mindreading based on the false belief task, are often adept at identifying knowledgeable individuals when contrasted with ignorant or inaccurate sources (e.g., Birch et al. 2008; Jaswal & Neely 2006; Koenig & Harris 2005; Sabbagh & Baldwin 2001). By age 4, children are also capable of drawing inferences about another person's knowledge that go beyond familiar associations to reflect an understanding of the underlying principles that make up a person's expertise (e.g., Lutz & Keil 2002). This suggests that young children understand both the contents and limits of another person's knowledge.

Conversely, young children's ability to assess their own knowledge accurately – a form of metacognition – is quite weak (Flavell et al. 1970; Mills & Keil 2004). For instance, 4- to 5-year-old children are notoriously poor at realizing how much they have learned of a new piece of information (Esbensen et al. 1997; Taylor et al. 1994). The younger they are, the more children overestimate their understanding (and underestimate their ignorance) of familiar objects and procedures, often not appropriately assessing their own level of knowledge until age 9 or later (Mills & Keil 2004). Thus, despite the evidence that children accurately judge what other people know, there seems to be a real developmental gap in applying these principles to their own knowledge.

One obstacle in applying this research to determining whether metacognition or mindreading is mastered first is that the kinds of questions used to ask children to reflect on their own knowledge are very different from the questions used to measure children's understanding of others' knowledge. Research examining how children think about their own knowledge often requires them to evaluate their own knowledge level using a scale or to estimate the number of problems they can answer correctly. In contrast, studies examining how children think about the knowledge of others typically require them to choose the source that would be able to give the most helpful or accurate information. Given that little research has been conducted directly comparing children's accuracy estimating the knowledge of others to their own knowledge, many open questions remain.

Therefore, in order to provide direct evidence in favor of or against Carruthers' hypothesis, we propose that additional research is necessary. Comparing the development of an understanding of one's own knowledge with an understanding of others' knowledge, or making self/other comparisons based on the other subcomponent skills required for mindreading and metacognition besides false belief, can provide information regarding the developmental trajectory for these skills. In designing this research, it is essential to identify analogous subcomponent skills important for mindreading and metacognition and test them using parallel measures. For instance, one could compare children's ability to identify what they know about a novel object versus another person's knowledge, given the same exact experience with the object. We believe such developmental evidence could go a long way in resolving this debate, and that, based on the research so far, it is likely to support Carruthers' hypothesis.

To conclude, there may be times when developmental evidence is of little use in resolving philosophical debates, such as when the debate is over the existence of free will or whether God exists. We do not believe this is one of those times; rather, Carruthers' proposal addresses a debate that is perhaps *best* understood in terms of child development. Certainly his hypothesis that mindreading is essential for metacognition poses specific empirical questions that can be tested using developmental methods. We feel strongly that careful research identifying analogous subcomponent skills necessary for mindreading and metacognition and examining their developmental trajectory will provide valuable evidence of whether mindreading really, truly is prior.

Varieties of self-explanation

doi:10.1017/S0140525X09000740

Dominic Murphy

Unit for History and Philosophy of Science, University of Sydney, Carslaw FO7, Camperdown, NSW 2006, Australia

d.murphy@usyd.edu.au

http://www.usyd.edu.au/hps/staff/academic/Dominic_Murphy.shtml

Abstract: Carruthers is right to reject the idea of a dedicated piece of cognitive architecture with the exclusive job of reading our own minds. But his mistake is in trying to explain introspection in terms of any one mindreading system. We understand ourselves in many different ways via many systems.

I agree with Carruthers that there is no piece of human cognitive architecture dedicated to introspection. But the right response is to abandon the search for one introspective metacognitive system, whether dedicated or coopted. We become aware of our states of mind by a variety of methods, which depend on a variety of systems.

Consider Hurlburt's experiments. Carruthers suggests that subjects employ their mindreading systems on their own behavior and circumstances when beeped. But he worries that not all cases can be handled like this, since there is nothing going on at the time of the beep that looks like evidence for one interpretation or another. To save the self-mindreading view, Carruthers is forced to conclude that subjects are basing their interpretation on their immediately prior behavior. This is a rather desperate expedient. People do not seem to report any awareness in these cases of their own behaviour or even, very often, their own states of mind. This supports the view that we put ourselves into a position to assert a state of mind by doing whatever we do to get into that state in the first place (Evans 1982).

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

When Hurlburt (1997) discusses the case of Donald, for instance, he does not show that Donald was interpreting himself at all when he was beeped. Donald noted, for instance, that his son had left the record player on again (Hurlburt 1997, p. 944). He reported attending to a fact about his environment. He was later brought to see the facts he noticed as evidence that he harbored unacknowledged anger toward his son. Donald went over his own transcript after beeping and interpreted his behaviour as he might anyone else's. This case helps Carruthers in one way, but it also shows something he misses about the beep cases. Donald did not report anything about his own mind or engage in any self-interpretation when he was beeped.

Many of Hurlburt's subjects do not, when beeped, report anything that looks like introspection or any other self-examination. They report thinking about the world, not about themselves. Carruthers mentions a subject who is wondering what her friend will be driving later; this is not a thought about one's own state of mind at all, and it is not clear why we need to call this interpretation. The subject is not self-interpreting. She is in a first-order state of wondering about the world, and being in that state is what lets her express it; she knows what she thinks by looking at the world, not by treating her deliberations as evidence for self-mindreading.

The assumption that ordinary deliberation must be accompanied by mindreading for one to report it is unnecessary and it gets Carruthers into trouble with those cases where subjects lack interpretative evidence from their own behaviour. Subjects report what they are thinking about, and often it is not themselves. If they were really engaged in self-mindreading, you would expect them to talk explicitly about their own beliefs and desires. But they often don't, which suggests that we should understand them as doing something other than either self-monitoring or self-interpretation. The absence of evidence for interpretation that Carruthers frets about is real, but it doesn't support the view he opposes. Rather, it supports the view that that we often know what we think by thinking about the world and not about ourselves. This supports the picture of self-attribution of belief that we find in Evans (1982, pp. 223–26), in which it is often just a matter of making up one's mind. Carruthers acknowledges that in cases of settled belief we can access our beliefs through memory. He reads Evans as showing that metacognitive access can arise through turning our mindreading capacities on our memory reports. But that is needlessly baroque. The simplest theory is that belief self-attributions are often just episodes of remembering.

I can assert a belief *that p* via the same procedure that would I go through in order to assert *that p*. This might be the result of working out what I believe via self-mindreading. But in other cases, when I state my belief *that p*, I am just remembering *that p* is true. Interpretative evidence is not needed. We do not have to assume that any interpretation is going on at all. In other cases, I put myself in a position to assert a belief by wondering if it is true. This is the way to handle the case of the woman who wonders what car she will go home in; we do not need to think of her as interpreting herself at all. Rather, the beeper leads her to say where her attention is focused, and it is focused on the world, so that is what she talks about. And attending to the world is not introspection or self-interpretation, even if it lets you say what you are thinking about.

Carruthers remains in needless thrall to the idea that metacognition needs a device that is directed at the mind. But when you self-attribute a propositional attitude, you are often not using an inward glance but an outward one: you are thinking about the world. Evans is concerned with this wider capacity to figure things out. When he talks of putting oneself in a position to report a belief, he is thinking of our abilities to deliberate about the world. Carruthers has isolated one way in which we may think about objects in the world – that is, we may treat them as things with minds, and we may look at our own behavior in that light too. But there is no reason to suppose that all our self-attributions come from self-mindreading. Introspection does not rely on any one system, neither an inner eye nor a mindreading device; it depends on all the ways one might think of states of affairs as believable or desirable. We can know our thoughts by looking at the world.

Global broadcasting and self-interpretation

doi:10.1017/S0140525X09000752

David Pereplyotchik

Department of Philosophy, Baruch College, City University of New York, New York, NY 10010.

res.cogitans@gmail.com

Abstract: Carruthers claims that global workspace theory implies that sensory states, unlike propositional attitudes, are introspectible in a non-interpretative fashion. I argue that this claim is false, and defend a strong version of the "mindreading is prior" model of first-person access, according to which the self-ascription of all mental states, both propositional and sensory, is interpretative.

According to the strong version of the "mindreading is prior" model (MPM), *all* metacognition is interpretative. On the weak version, we have non-interpretative access to both sensory states and propositional attitudes. Carruthers' version of MPM is a middle-ground position. In accord with the strong version, he insists that first-person awareness of propositional attitudes is always self-interpretative and, hence, never "introspective" (in his quasi-pejorative sense). However, he denies that self-attribution of "sensory-imagistic" states is interpretative, claiming that such states are introspectively available as *data* for the mindreading system. On his view, "the mindreading system can receive as input any sensory or quasi-sensory (e.g., imagistic or somatosensory) state that gets 'globally broadcast' to all [cognitive] systems" (sect. 2, para. 3). The set of such states includes "perceptions, visual and auditory imagery (including sentences rehearsed in 'inner speech'), patterns of attention, and emotional feelings" (sect. 2, para. 6).

I argue that Carruthers' appeal to the distinction between sensory states and propositional attitudes involves an error, and that avoiding this error leads to the collapse of his view into one of the competing versions. The preferable collapse is, I argue, toward the stronger view.

The states that Carruthers takes to be introspectively available are supposed to be "sensory-imagistic," not conceptualized or propositional. But it is not obvious that perceptual judgments satisfy this description. Perceptual judgments are plainly a species of propositional attitude. True, such states may also have qualitative properties, but they are nevertheless constituted by concepts, and have "sentence-sized" intentional contents. Carruthers acknowledges this in discussing how the mindreading faculty would make use of perceptual judgments:

> Receiving as input a visual representation of a man bending over, for example, [the mind-reading system] should be capable of forming the judgment, "I am seeing a man bending over." (*At least, this should be possible provided the visual state in question has been partially conceptualized by other mental faculties, coming to the mindreading system with the concepts* man *and* bending over *already attached*). (target article, sect. 2, para. 4, my emphasis)

In explaining why perceptual judgments appear on his list of introspectible states, Carruthers's appeals to global workspace theory. That perceptual judgments are introspectible is, in his view, "pretty much mandatory once we buy into a global broadcasting architecture" (sect. 2, para. 8). That's because, in perception, "the initial outputs of the visual system interact with a variety of conceptual systems that deploy and manipulate

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

perceptual templates, attempting to achieve a 'best match' with the incoming data. . . . [T]he result is globally broadcast as part of the perceptual state itself' (sect. 2, para. 10).

Although this is certainly plausible, Carruthers neglects the fact that similar grounds are available for the claim that other propositional attitudes are broadcast as well. Indeed, given his concession that global workspace theories allow for the broadcasting of at least one propositional attitude, one wonders why he assumes that such theories would *not* allow for the broadcasting of all the rest. On the face of it, the claim that *all* propositional attitudes can be globally broadcast has much going for it. Intentions, for instance, routinely recruit a wide array of cognitive resources, as do the conceptual-intentional aspects of emotions like fear and anger (e.g., *that* one is being attacked). Why not count these as instances of global broadcasting? Carruthers does not say.

Pending further argument, we should assume, *pace* Carruthers, that global workspace theory *does* allow for the broadcasting of all propositional attitudes. If so, then whatever we say about first-person access to sensory states, we should say the same about first-person access to propositional attitudes.

Do these considerations support the view that the mindreading system has direct, non-interpretative access to *all* mental states, both propositional and sensory? Not if one also rejects Carruthers' assumption that globally broadcast states are ipso facto available to the mindreading system in a *non-interpretative* fashion. Below, I explore grounds for adopting the strong version of MPM, according to which self-attribution is interpretative in the case of all mental states.

Interpretation takes place by deploying a propositional attitude that emerges from a background of theoretical commitments. Consequently, the cost of embarking on an interpretative venture is the possibility of partial misconstrual or wholesale error. These characteristics of interpretative activity fit well with Carruthers' usage of the term "interpretative," as applied to mechanisms of self-attribution.

As Rosenthal (2005) has argued, self-attribution is a matter of tokening potentially erroneous, theoretically loaded propositional attitudes – occurrent higher-order thoughts (HOTs). On this view, confabulation and error occur even with regard to sensory states. Dental fear, for instance, is a phenomenon in which dental patients under the drill mistake fear, anxiety, and a sensation of vibration for pain in a fully anaesthetized or nerveless tooth – a compelling demonstration that HOTs need not be veridical.

Nevertheless, judged on independent grounds, self-attributions of sensory states are often relatively accurate. Doubtless, this consideration compels theorists to posit a reliable monitoring mechanism, such as Carruthers' mindreading system. But, as Rosenthal points out, simply positing such a mechanism amounts to no more than *stipulating* a solution to the problem of explaining the frequent accuracy of HOTs about sensory states. An explanatory account of the mechanism's accuracy is not provided.

Extending Sellars's (1956/1997) treatment, Rosenthal argues that HOTs concerning sensory states arise as a result of a creature's reflection on cases in which its perceptual judgments are mistaken. The creature formulates a rudimentary theory, in effect positing qualitative sensory states as the causes of nonveridical perceptual judgments. Against the background of such a theory, the creature is disposed, for instance, to construe itself as having a sensation of red when perceiving a red object.

Carruthers gives no grounds for rejecting this alternative and appealing picture. Global broadcast theory does not, by itself, settle the issue, for it is consistent with the claim that the mindreading system relies on a tacit theory in interpretatively self-ascribing sensory states. Nor does the data from autistic children disconfirm Rosenthal's view, which allows that even nonlinguistic, cognitively unsophisticated creatures may come to have HOTs concerning their sensory states. All that is required is that such creatures take note of their perceptual errors and account for them.

ACKNOWLEDGMENT

I am grateful to David Rosenthal for helpful comments on an earlier draft.

Introspection and interpretation: Dichotomy or continuum?

doi:10.1017/S0140525X09000764

Richard E. Petty[a] and Pablo Briñol[b]

[a]*Department of Psychology, Ohio State University, Columbus, OH 43210;* [b]*Departamento de Psicologia Social, Universidad Autonoma de Madrid, 28049 Madrid, Spain.*

petty.1@osu.edu www.psy.ohio-state.edu/petty
pablo.brinol@uam.es www.psy.ohio-state.edu/gap/Pablo/pablo.htm

Abstract: Judgments vary in the extent to which they are based on interpretation versus relatively direct access to mental contents. That is, a judgment might require a trivial amount of interpretation (e.g., translating one's immediately accessible "inner speech") or a rather substantial amount of confabulation. Recognizing this continuum of interpretation underlying judgment could be more fruitful than debating a categorical introspection versus interpretation distinction.

Some prior authors have noted that people have no unique access to *why* they believe what they believe (e.g., Nisbett & Wilson 1977). Others have gone a step further and postulated that people do not know their own attitudes (e.g., I like ice-cream) but must construct them when needed from other available information that they either retrieve from memory (e.g., ice cream tastes good) or extract from the immediate context (e.g., Schwarz & Bohner 2000). Carruthers takes this "constructivist" position to the ultimate extreme by arguing that people have no direct access to any attitudes *or* relevant beliefs. According to this view, introspection does not exist, and is merely an illusion. Furthermore, he provides many examples where people either clearly or plausibly are confabulating when they express what they believe. In his view, at best individuals only know what they feel and perceive, not what they think. In our view, it is not clear why an intelligent organism would have evolved to have direct access to its feelings and perceptions but not its cognitions.

Nevertheless, Carruthers has an important point. Whenever someone expresses a belief or has a thought, some degree of interpretation likely is involved, if only to understand the meaning of the "inner speech" in which the thought is expressed. Thus, if a person has a positive reaction to some stimulus (ice-cream), this can be translated into "it's good" (Fazio 1985). Or even if the word "good" immediately springs to mind, the meaning of the word must be understood by the self if an internal thought, or by others if expressed. However, this very minimal form of "interpretation" is very different from the kind of interpretation involved in most of the examples of confabulation provided by Carruthers. Indeed, we argue that it may not be wise to think of introspection and interpretation as dichotomous categories in which to place any given judgment. Rather, there are various degrees of interpretation. At the low end of the interpretation continuum, judgments are introspection-like in that they involve at most some trivial interpretation. At the other end of the continuum, the judgment is totally confabulated from external sources also available to outside observers.

Although dichotomous categories can be useful in understanding some phenomena, as illustrated by the distinction between primary and secondary (meta-) cognition, we believe that it is not conducive to understanding human information processing

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

to simply lump all judgments into the same overarching "interpretation" category and stop there. This is because putting all judgments into the same category might suggest that there are no meaningful differences *within* the catetory. In contrast to lumping all judgments together into one interpretation category, we espouse a continuum view in which people express beliefs based on very little interpretation in some cases but based on extensive confabulation in others. We further argue that differences in the *degree* of interpretation are meaningful.

Previous research on psychological elaboration provides one instance of the usefulness of the continuum view. The term "elaboration" is used in social psychology to describe that people add something of their own to the specific information provided, for example, in a persuasive communication. In the domain of social judgment, variations in elaboration are consequential. For example, when people are relatively unmotivated or unable to think, they are more likely to rely on immediately accessible information that originates either internally (one's attitude) or externally (e.g., the attractiveness of the message source). When people are more motivated and able to think, then these initial reactions and the judgments that follow from them can be overridden by more complete interpretative analyses. Furthermore, judgments based on high levels of elaboration are more consequential than those based on low levels (Petty & Cacioppo 1986).

Viewing interpretation as a continuum has a number of implications. Most obviously, it means that interpretation can go from zero (i.e., introspection) to extensive. More interestingly, the continuum view suggests that the point on the continuum that corresponds to minimal or trivial interpretation has more in common with zero interpretation than it does with extensive interpretation. One can draw an analogy to a distinction that attitude theorists used to favor between attitude formation versus attitude change. Different mechanisms of influence were thought to be operative depending on whether a person had an existing attitude or did not (a categorical view). Today, it is more common to think of attitudes as falling along a continuum such that they can vary in how accessible they are or upon how much knowledge they are based. An attitude formation situation would be present when a person has no prior attitude. But, a continuum approach to attitudes suggests that a person who has an attitude that is difficult to bring to mind and based on little information (Person B) has more in common with an individual who has no attitude (Person A) than a person who has an attitude that comes to mind spontaneously and is based on much knowledge (Person C). That is, the first two individuals – A and B – are more similar to each other in psychologically relevant ways than they are to C, despite the fact that a dichotomous approach places Person B in a different category from A and in the same category as C. So too is it the case that a judgment based on minimal interpretation (B) is closer to a judgment based on no interpretation (A) than it is to a judgment based on extensive interpretation (C; see Fig. 1).

In sum, we conclude that an all-or-none frame regarding the existence of introspection may not be the best way to make the most progress in understanding social judgment. Instead, drawing from the literature on elaboration and attitude strength, we suggest that it might be more fruitful to approach interpretation as a continuum where the low end is anchored at introspection. That is, sometimes interpretation can be quite minimal as when people recall their birth-date or liking of a favorite book. At other times, interpretation can be quite extensive, such as when there is either nothing relevant to recall or one's interpretation totally overwhelms any mental content introspected.

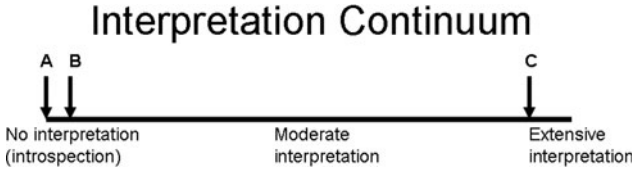


Figure 1 (Petty & Briñol). Continuum of extent of interpretation underlying judgment

Overlooking metacognitive experience

doi:10.1017/S0140525X09000776

Joëlle Proust

Department of Cognitive Studies, Ecole Normale Supérieure, and Institut Jean-Nicod, EHESS and ENS, 75005 Paris, France.

jproust@ehess.fr

http://joelleproust.hautetfort.com

Abstract: Peter Carruthers correctly claims that metacognition in humans may involve self-directed interpretations (i.e., may use the conceptual interpretative resources of mindreading). He fails to show, however, that metacognition cannot rely exclusively on subjective experience. Focusing on self-directed mindreading can only bypass evolutionary considerations and obscure important functional differences.

Carruthers' main goal is to show that metacognition is a form of self-directed interpretation, akin to other-directed mindreading. Introspection, he claims, defined as "any reliable method for forming beliefs about one's own mental states that is *not* self-interpretative and that differs in *kind* from the ways in which we form beliefs about the mental states of other people" (sect. 1.4, para. 3, emphasis in the original), is not needed to have access to one's own mental attitudes. One can agree with the author that metacognition in humans *may* involve self-directed interpretations (i.e., *may* use the conceptual interpretative resources of mindreading), without accepting the stronger claim that metacognition can *never* be based on "introspection."

In cognitive science, "metacognition" refers to the capacity through which a subject can evaluate the feasibility or completion of a given mental goal (such as learning a maze, or discriminating a signal) in a given episode (Koriat et al. 2006). In Carruthers' use, however, metacognition *refers to* first-person metarepresentation of one's own mental states; as a result, the theoretical possibility that metacognition might operate in a *different* representational format cannot be raised (Proust, in press b). Revising the meaning of a functional term such as "metacognition" is a bold strategy. It generally seems more adequate to leave it an open empirical matter whether a capacity of type A (reading one's own mind) or type B (evaluating one's cognitive dispositions) is engaged in a particular task. A revision is deemed necessary, according to Carruthers, because "B" capacities in fact always involve self-directed mindreading; therefore apparent contrary cases (self-evaluation in non-mindreading animals) either (1) are simply instances of first-order types of learning, and/or (2) are capacities "too weak to be of any interest" (Carruthers 2008b, p. 62; cf. target article, sects. 5.1 and 9).

Two methodological problems, however, hamper the discussion so conceived. First, it is quite plausible that, in *human* forms of metacognition (as instantiated in speech production, metamemory, etc.), judgments of self-attribution *redescribe* elements of metacognitive experience. Metacognitive feelings might, on this view, represent subjective uncertainty and guide noetic decision-making, without needing to involve a conceptual interpretative process. What needs to be discussed, in order to establish the superiority of model 4, is whether or not subjects can rely on dedicated feelings alone to monitor their ongoing cognitive activity.

A second, related problem is that Carruthers' discussion conflates two domains of self-control, namely, the control of one's

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

physical actions through perceptual feedback and the control of one's mental actions through metacognitive feedback (see sects. 5.1 and 9). Meta-action, however, is only functionally similar to metacognition when a metarepresentational reading is imposed on both, in spite of their different evolutionary profiles (Metcalfe & Greene 2007; Proust, in press a). If extracting, from a given task context, an evaluation of the mental resources available to complete the task were just another case of first-order action control, then one might agree that B-metacognition is nothing other than executive function. But metacognitive and executive abilities can be dissociated in schizophrenia (Koren et al. 2006). Mental action control is thus distinct both from executive memory as usually understood and from physical action control.

These methodological problems strongly bias the discussion against models 1 and 3. Here are three examples.

1. Our metacognitive interventions don't require introspection; they have no direct impact on cognitive processing (sect. 5.1).

From a B-sense viewpoint, prediction and evaluation of one's mental states and events presuppose appreciating one's subjective uncertainty regarding correction, adequacy, and so on, of first-order decisions or judgments; this evaluation does not require that the target states are *represented qua mental*. For example, a child chooses to perform one memorization task rather than another by relying not on *what she knows about memory*, as the author claims, but on the *feeling* she has that one task is easier than another (Koriat 2000; Proust 2007). The impact on decision is quite direct, and it is independent of mindreading.

2. A combination of first-order attitudes is sufficient to explain how animals select the uncertainty key in Smith et al.'s metaperceptual paradigm (sect. 5.2).

If this is correct, how can monkeys rationally decide to opt out when no reinforcement of the uncertainty key is offered, and when, in addition, novel test stimuli are used? Why should there be transfer of the degree of belief associated with first-order items to novel tasks where these items are no longer included? A second rule must apply, as Carruthers (2008b) himself admits: having conflicting impulsions to act or not to act on a given stimulus, the subject becomes uncertain of its ability, say, to categorize. So the decision to act depends, after all, on subjective – not objective – features. Can these subjective features influence behavior *only* through their being metarepresented ? This is the crucial question that fails to be raised.

3. "Evidence suggests that if mindreading is damaged, then so too will be metacognition" (sect. 10, para. 10).

Clinical research on autism and on schizophrenia suggests rather a dissociation of metacognitive and mindreading skills as predicted by model 1 (cf. Bacon et al. 2001; Farrant et al. 1999; Koren et al. 2006). However, its relevance for the present discussion is downplayed as a smart behaviorist effect; introspection in patients with autism is rejected because it is not "metacognitive in the right sort of way" (sect. 10, para. 5). Negative results in meta-action from patients with autism are presented as evidence for impaired metacognition. Such appraisals implicitly appeal to the preferred metarepresentational interpretation of metacognition under discussion. Similarly, rejecting the relevance of metacognitive capacities which are "too weak to be of any interest" presupposes recognizing the superiority of model 4 over models 1 and 3.

A fair examination of the contribution of "introspection" in metacognition in models 1 and 3 would require studying the respective roles of control and monitoring in nonhuman and human epistemic decisions, in experimental tasks engaging metaperception, metamemory, and planning (Koriat et al. 2006; Proust, in press a). Focusing on self-directed mindreading can only bypass evolutionary considerations and obscure important functional differences.

ACKNOWLEDGMENTS

I am grateful to Dick Carter, Chris Viger, and Anna Loussouarn for their remarks and linguistic help. The preparation of this commentary was supported by the European Science Foundation EUROCORES Programme CNCC, with funds from CNRS and the EC Sixth Framework Programme under Contract no. ERAS-CT-2003–980409.

Guilt by dissociation: Why mindreading may not be prior to metacognition after all

doi:10.1017/S0140525X09000788

Philip Robbins

Department of Philosophy, University of Missouri–Columbia, Columbia, MO 65211.

robbinsp@missouri.edu

http://philosophy.missouri.edu/people/robbins.html

Abstract: Carruthers argues that there is no developmental or clinical evidence that metacognition is dissociable from mindreading, and hence there is no reason to think that metacognition is prior to mindreading. A closer look at the evidence, however, reveals that these conclusions are premature at best.

In psychology, evidence of dissociation comes in one of two forms: synchronic or diachronic. Two capacities are synchronically dissociable if there are adults in whom the first capacity is defective and the second capacity is intact, or vice versa. Evidence for dissociation of this sort comes from studies of clinical mental disorders, such as autism and schizophrenia. Two capacities are diachronically dissociable if the first capacity emerges before the second capacity does, or vice versa. Evidence for dissociation of this sort comes from studies of normally developing children. A central issue in the target article is the dissociability – or lack thereof – of metacognition (first-person metarepresentation) from mindreading (third-person metarepresentation). After waving the issue of diachronic dissociation mostly to one side (sect. 4), Carruthers argues that there is no good evidence of synchronic dissociation (sects. 9 and 10). With respect to the first point, his dismissal of the developmental evidence is premature; on the second point, he makes a slightly better case, but the evidence isn't all in – so the jury is still out. Or so I argue.

Let's start with the case of synchronic dissociation and the view from developmental psychology. Carruthers opens his discussion here by citing a large-scale meta-analysis that found no evidence of developmental self-other asymmetries on metarepresentation tasks (Wellman et al. 2001). This is grist for the "mindreading is prior" mill but only up to a point. That's because the meta-analysis by Wellman et al. was exclusively concerned with studies using standard false-belief tasks and no other measures of metarepresentational capacity. And once we take into account the results of studies using such nonstandard measures, the idea that metacognition is developmentally prior to mindreading becomes more plausible (Nichols & Stich 2003; see also Robbins 2006).

A key piece of evidence here comes from a study by Wimmer et al. (1988) using a version of the "seeing leads to knowing" paradigm. (Curiously, Carruthers never cites this study, though he does cite a later study of children with autism that employed much the same paradigm, albeit with different – and for his purposes, more favorable – results; see Kazak et al. 1997.) In this study, done with normally developing 3-year-olds, subjects were divided into two groups. In the first group, the children were instructed to look inside a box; in the second group, visual access to the interior of the box was denied. Subjects in both groups were then asked whether or not they knew what was in the box, and most of the children in each group gave the correct answer: those in the first group said yes, those in

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

the second group said no. Subsequently, children in both groups observed another person either looking inside of the box or not looking into it. They were then asked whether or not the person they had observed knew what was inside the box. The answers they returned were surprising (especially for anyone familiar with the literature on egocentric biases). For example, 14 children represented their own epistemic state correctly while misrepresenting the epistemic state of the other person, whereas only 2 children displayed the opposite pattern of responses (i.e., correct for other, incorrect for self) – a highly significant contrast ($p < .01$). In short, results from the Wimmer et al. study suggest that young children are better at reporting their own knowledge state than the knowledge state of others with whom they knowingly share access to the relevant information. Because knowledge is a propositional attitude par excellence, this looks like good support for the synchronic dissociability of metacognition from mindreading in precisely the domain of interest, namely, the metarepresentation of propositional attitudes.

Let's turn now to the issue of synchronic dissociation. Carruthers points out that passivity symptoms in schizophrenia, such as thought insertion and delusions of control, need not be (and probably are not) due to a breakdown in metacognition, and hence that studies of passivity-symptomatic schizophrenia – a condition in which mindreading appears to be relatively intact – do not support the idea that metacognition is defective in this condition. But this is a double-edged sword. For if the "mindreading is prior" view is correct, then metacognition should be impaired in those subtypes of schizophrenia in which mindreading is defective. Passivity-symptomatic schizophrenia is unrepresentative in this respect, in that patients with this subtype of the disorder tend to perform normally on standard first- and second-order false-belief tasks (Corcoran 2000). By contrast, individuals in the paranoid subtype of schizophrenia perform poorly on a wide range of mindreading tasks, including tasks involving the attribution of intentions and the understanding of jokes, hints, and conversational implicatures (Brüne 2005). Indeed, the defining symptoms of paranoid schizophrenia include persecutory delusions and delusions of self-reference, both of which involve misattributing to other people intentions toward oneself. Exactly what explains this tendency is a matter of controversy, but its characterization in terms of a deficit in the mindreading system is not (Blackwood et al. 2001). Hence, if the "mindreading is prior" view is right, then we should expect to find impaired metacognition in these patients, especially in the domain of intention attribution; otherwise, we should conclude that metacognition and mindreading are dissociable after all. As things stand, however, there appears to be no credible evidence to this effect.

There are at least two possible explanations for this fact, only one of which is immediately damaging to Carruthers' position. The first possibility is that evidence of metacognitive impairments in paranoid schizophrenia is lacking because metacognition is spared in this disorder. That would be bad news for the "mindreading is prior" view. A second possibility, however, is that evidence of this sort is lacking simply because the requisite empirical studies have not been done yet. This second possibility seems at least as likely as the first, especially as far as the metacognition of propositional attitudes like intention is concerned. Schizophrenia researchers have paid more attention to the metacognition of emotions, that is, nonattitudinal mental states. But here the news isn't good for Carruthers either, for it appears that third-person deficits, such as difficulties with face-based emotion recognition, need not be accompanied by first-person deficits, such as difficulties with recognizing one's own emotions (Brunet-Gouet & Decety 2006). It is entirely possible, then, that future studies of the metacognition of intention in schizophrenia will point toward a similar dissociation. And in that case, proponents of the "mindreading is prior" view will have some explaining to do.

Social-affective origins of mindreading and metacognition

doi:10.1017/S0140525X0900079X

Philippe Rochat

Department of Psychology, Emory University, Atlanta, GA 30322.
psypr@emory.edu

Abstract: The engineer's look at how the mind works omits a central piece of the puzzle. It ignores the dynamic of motivations and the social context in which mindreading and metacognition evolved and developed in the first place. Mindreading and metacognition derive from a primacy of affective mindreading and meta-affectivity (e.g., secondary emotions such as shame or pride), both co-emerging in early development.

William James in his 1872 publication entitled "Are We Automata?" makes the point that "a succession of feelings is not one and the same thing with a feeling of succession, but a wholly different thing" (James 1872, p. 5). James insists that: "The latter feeling requires a self-transcendency of each item, so that each not only is in relation, but *knows* its relation, to the other." James concludes that this "self-transcendency of data" constitutes what consciousness is, in his own words "what constitutes the conscious form" (p. 6). It is also what makes us nonreducible to automata.

If we agree with James that we are more than machines, an engineer look at self-consciousness (e.g., metacognition and introspection) and the consciousness of others (mindreading and theories of mind) is insufficient and does not capture the essence of the phenomena.

At the core of these issues, there is more than the rational attribution of mental states to the self or to others (theories of mind). There are also *feelings* and *emotions*, what make us sentient and incommensurate to machines. In the social domain, "self-transcendency" corresponds first and foremost to the *feeling* of our relation to others, not just to a propositional attitude turned toward others or onto the self (metacognition).

Children in their social-affective development tell us that upstream to mindreading and metacognizing, there is the affective urge to see the self as seen and evaluated by others – the deep-seated affective propensity to have "others in mind" (Rochat 2009).

The origins of a tendency to attribute attitudes (i.e., adopt a propositional attitude stance) toward others as well as toward the self are rooted in the basic propensity to know *where we stand emotionally in relation to others*: whether we feel affective distance or closeness, a sense of affiliation and approval, or on the contrary a sense of disapproval and rejection from others.

There is now much evidence showing that long before children display signs of explicit "theorizing" about the mental states of others, as well as their own, whether for example they have an appreciation of their own knowledge or ignorance, they have already developed astute implicit intuitions of whether they can more or less "trust" someone, or "trust" themselves in what they believe and know.

By 2 months, infants expect someone to behave in a certain way in certain social contexts. They show distress and emotional withdrawal toward an adult who abruptly and for no apparent reasons adopts a frozen facial expression. They also show resistance in recovering positive affects following a still-face episode (Tronick 2005; Tronick et al. 1978). Infants are indeed highly sensitive to the attention received from others, constantly gauging how much others are engaged toward them (Reddy 2003). They are also quick to develop a preference for the interactive contingency style of their mother, generalizing such preference when encountering strangers (Bigelow & Rochat 2006).

These facts demonstrate that at the roots of mindreading and metacognition, there is a complex pragmatics of intentional communicative exchanges that is not just rational, but ontologically affective and emotional. The first nonverbal language of children is well documented, expressed from at least 2 months of age in

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

the context of face-to-face interactions, made of mutual gazes, precise contingency of exchanges and turn taking, with compulsive "motherese" and other affective markers from adults.

The syntax of this early nonverbal grammar (e.g., when I smile, the other should not scream or cry) and the semantics that children derive from it (e.g., if I smile and the other looks away, something is wrong) is primarily emotional and affective. These rich and reciprocal communicative exchanges quickly determine a proto-rationality that is expressed in the epistemic as well as affective trust that children from the earliest age place in others, as well as eventually into themselves.

Topal et al. (2008) provide further evidence of the deep affective and emotional roots of mindreading and metacognition. These authors show that the famous Piagetian stage 4A-not-B error of object permanence found in 10-month-olds and younger infants, depends in part on the presence or absence of subtle communicative and emotional cues (i.e., eye contact, motherese, affective attunement, or social contingency) from the experimenter hiding the object. Infants are inclined to interpret the adult as *wanting* to teach them something (a hiding game), while having a good, intimate social time of sustained mutual attention. Topal et al. show that 10-month-olds' perseverative search errors can be induced by pragmatic misinterpretation of an adult's intentions.

By the time children objectify themselves in mirrors passing the famous mark test (Amsterdam 1972), they also begin to express secondary emotions such as embarrassment, shame, or pride (Lewis 1992). Early on, infants demonstrate a complex appreciation about their public appearance, long before they express rational theories of mind and metacognition. Affectively, children first feel how others like or dislike them (affective reading), gauging also how they think they might be liked or disliked by others (meta-affectivity). Both seem to co-emerge in development, necessarily codependent.

From a developmental vantage point, affective reading and meta-affectivity are ontologically linked, representing two sides of the same coin. Findings on early social and emotional development demonstrate that there is not much empirical grounding for conceptualizing them as separate processes.

In short, let us not forget that we are born in need of social attention; and it is, I would suggest, in this primary motivational context that early social expectations develop. These expectations, co-constructed in the history of interactions with others, become by 4 to 5 years of age explicit "theories" about self and others' propositional attitudes. This primary motivational context represents an invaluable source of information that cannot be dismissed or overlooked, particularly if one attempts at capturing "how we know our own mind."

Dealing with the question outside of any motivational context, as done by Carruthers in the target article, is a disembodied, abstract exercise. By analogy, it is like playing chess or moving armies on a map instead of being physically and emotionally engaged on the battlefield. The battlefield is riskier, but closer to what it is all about.

Metacognition, mindreading, and insight in schizophrenia

doi:10.1017/S0140525X09000806

Ben Wiffen and Anthony David

Department of Psychological Medicine and Psychiatry, Institute of Psychiatry, King's College London, London SE5 8AF, United Kingdom.

ben.wiffen@iop.kcl.ac.uk a.david@iop.kcl.ac.uk

www.iop.kcl.ac.uk/staff/?go=10055

Abstract: Mindreading in schizophrenia has been shown to be impaired in a multitude of studies. Furthermore, there is increasing evidence to suggest that metacognition is damaged as well. Lack of insight, or the inability to recognise one's own disorder, is an example of such a failure. We suggest that mindreading and metacognition are linked, but separable.

Here we review the evidence for deficits in mindreading (or Theory of Mind [ToM]) tasks in schizophrenia and look at some work on true metacognitive tasks in schizophrenia, in which schizophrenia patients also display deficits. We argue that what psychiatrists refer to as a "lack of insight" is an example of a failure to make judgements of themselves that they do make of others.

Since the publication of Chris Frith's seminal *Cognitive Neuropsychology of Schizophrenia* (Frith 1992), research investigating ToM in schizophrenia has been widespread. A deficit in mindreading is clearly demonstrable in schizophrenia, but whether this is specific and causally related to certain symptoms or merely further evidence of generalised cognitive impairment found in virtually all patients with the disorder remains unclear (Harrington et al. 2005). Sprong et al.'s (2007) meta-analysis showed a large and statistically significant impairment on ToM tasks (Cohen's $d = 1.26$) across all schizophrenia sub-types and tasks, larger than studies restricted to "paranoid" patients, and claimed that there was support for mindreading impairment as a trait marker of schizophrenia.

Studies apparently supporting Carruthers' claim that mindreading is intact in "passivity" schizophrenia have had small sample sizes in their group (e.g., 7 in Corcoran et al. 1995; 1 in Pickup & Frith 2001). Furthermore, these studies may be confounded by other common symptoms; that is, someone with passivity symptoms may well exhibit paranoid or negative symptoms. So claiming that passivity symptom patients "perform normally when tested on batteries of mindreading tasks" (target article, sect. 9, para. 1) is overly bold. Indeed, recent work comparing controls with "passivity" patients showed different neural activity in ToM tasks, even though responses were broadly similar (Brüne et al. 2008). This suggests that a different cognitive strategy is used by such patients in mindreading tasks.

We question whether passivity "experiences" are ever pure metacognitive failures. Although classical phenomenologists like Kurt Schneider attempted to separate what in today's parlance we would call the experience from the attribution, the tiny corpus of examples in the literature invariably link the experience with a "psychotic" explanation – that a person or organisation is doing the movement or speaking through me using some dastardly device or technology. When the experience alone is reported, we can't be sure that there is no abnormal attribution. Our clinical impression is that the two are inextricably linked – at least in people presenting to clinical services. We believe this is the essence of Frith's model: that delusions, such as those of alien control, may build on passivity experiences but are the result of ToM reasoning which has gone awry. One might even say that psychosis is the result of *excessive* attribution – usually of malign intent – to people or things. People with psychosis have a theory of mind – it is just the wrong theory.

We would also argue that there is evidence for impairment in metacognition in schizophrenia in a truer sense than the "weak" variety described. The Beck Cognitive Insight Scale (Beck et al. 2004) shows differences between schizophrenia patients and controls in their agreement with statements on subscales assessing self-reflection and self-certainty (Warman et al. 2007). Furthermore, some studies show considerable deficits in psychotic patients' assessment of their own poor neurocognitive function (Medalia & Thysen 2008), although accurate assessment of such deficits may coexist with lack of awareness of the implausibility of beliefs suggesting fractionation of metacognitive awareness (Gilleen et al., in press). However, metacognitive failure is not for a want of trying. Bedford and David (2008) showed that patients actually place time and value in self-reflection, but just struggle to do it accurately. Additionally, patients' metacognitive performance on the Wisconsin Card Sort Task (i.e., the

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

confidence they place in the decisions they make) predicts disrupted insight in schizophrenia, suggesting a link between a failure of metacognition in a narrow sense, to a broader one (Koren et al. 2004).

What psychiatrists refer to as a lack of insight (the ability to judge oneself to have a mental disorder and to "relabel" abnormal experiences as related to such a disorder) is very common in schizophrenia and has even been conceptualised as a core aspect of the condition (David 2004). It seems to be an example of a failure of metacognition. Patients fail to make accurate judgements about themselves – in relation to their thoughts and experiences – and instead choose often bizarre and impossible explanations. However, it does not seem to be the case that patients who lack insight into their own illness are completely unable to make accurate judgements about others. Their mindreading abilities remain intact to the extent that they are relatively unimpaired in their attributions of "madness" to others. Rockeach's *Three Christs of Ypsilanti* (Rockeach 1964) reports the reactions of three patients presenting with broadly the same delusion. They all continued to believe in their identity whilst rejecting the others as mad: "Truth is truth, no matter if only one person speaks it" (p. 150), claims one, steadfastly defending his identity.

Indeed, empirical studies using vignettes have shown no differences between patients and controls in attribution of illness. Startup (1997) showed that patients could distinguish between normal and psychotic thoughts, feelings, and behaviours in the vignettes but this showed no correlation with the patients' insight into their own condition. McEvoy et al. (1993) found that patients correctly labelled characters with psychotic symptoms in vignettes as suffering from mental illness, but failed to note their own similarity to the character – although it was obvious to the patients' doctor. Further studies assessing the relationship of mindreading and metacognition as related to their applications in clinical insight are required: a simple correlation between the two functions seems unlikely (Pousa et al. 2008).

We note that metacognition can be enhanced by presenting the self as if another person – for example, by replaying a video of himself when ill to the patient (Davidoff et al. 1998). A conceptually similar approach taken to neurological patients with anosognosia for hemiplegia seems to yield increased awareness (Marcel et al. 2004). These observations are not decisive in dissociating mind reading from metacognition but suggest the following hypotheses: That *self*-awareness may at least make use of cognitive mechanisms which afford awareness of *others*' intentions and beliefs; that this is not necessarily automatic and may be prevented, presumably by attentional mechanisms or strategic failures (not looking), but can be overcome by adopting a third-person perspective to the self. Thus, some practical benefit may accrue from considering metacognition and ToM as linked but separable.

Metacognition may be *more* impaired than mindreading in autism

doi:10.1017/S0140525X09000818

David M. Williams,[a] Sophie E. Lind,[b] and Francesca Happé[c]

[a]*Institute of Child Health, University College London, London, WC1 N 1EH, United Kingdom;* [b]*Department of Psychology, City University, London, EC1V 0HB, United Kingdom;* [c]*Social, Genetic, and Developmental Psychiatry Research Centre, Institute of Psychiatry, London, SE5 8AF, United Kingdom.*

d.williams@ich.ucl.ac.uk Sophie.Lind.1@city.ac.uk
f.happe@iop.kcl.ac.uk

Abstract: This commentary focuses on evidence from autism concerning the relation between metacognition and mindreading. We support Carruthers' rejection of models 1 (independent systems) and 3 (metacognition before mindreading), and provide evidence to strengthen his critique. However, we also present evidence from autism that we believe supports model 2 (one mechanism, two modes of access) over model 4 (mindreading is prior).

Impaired metacognition in autism. We agree with Carruthers' claim that both mindreading *and* metacognition are impaired in autism, and that this speaks against models 1 and 3. However, we wish to provide more decisive evidence for impaired metacognition in autism, given that the evidence cited by Carruthers is problematic. For example, contrary to Carruthers' suggestion (sect. 10, para. 10), Kazak et al. (1997) did *not* find statistically significant differences between participants with and without autism in either mindreading or metacognition (see Kazak et al., p.1005).

Clearer evidence for metacognitive deficits among children with autism emerges from Williams and Happé (in press a), who assessed awareness of own intentions. Compared to age- and ability-matched comparison children, children with autism were significantly less likely to (a) correctly report their reflex movements as unintentional, and (b) correctly recognise their own mistaken actions (drawing a picture different to that intended, through experimenter manipulation) as unintended. The performance of children with autism on these measures was significantly associated with performance on traditional false belief tasks, independent of verbal ability. These findings suggest that children with autism have a limited ability to represent their own intentions and that these difficulties are fundamentally associated with established difficulties in representing others' mental states (in this case, false beliefs). These findings provide robust evidence against both models 1 and 3, but do not differentiate models 2 and 4.

Impaired episodic memory in autism implies impaired metacognition. Several researchers (e.g., McGeer 2004; Nichols & Stich 2003; Raffman 1999) have claimed that the autobiographical reports of individuals with autism show that metacognition is intact in autism (supporting models 1 and 3). However, none of these authors, nor Carruthers in his target article, distinguishes between semantic and episodic memory.

Episodic memory is associated with consciously remembering *personally experienced* events, whereas semantic memory is concerned with *factual* information. Caution must be exercised when attributing to people with autism memory processes of the *episodic* kind. I may *know* that a particular event has happened to me in the past, and hence report details of the event quite accurately, without actually *remembering* the event. Only this latter kind of "remembering" is thought to rely on metacognition (e.g., Perner 2000).

Contrary to models 1 and 3, it may be that people with autism do not engage the same metacognitive processes as typical individuals do when reporting events from their past (Lind & Bowler 2008). Lind and Bowler (under revised review; see also Lind 2008) found that children with autism ($n = 53$) were as able as age- and ability-matched comparison participants to distinguish events that had occurred from events that had not (whether or not a picture had been picked up and named earlier in the test session). However, participants with autism were significantly impaired at remembering the *source* of such events (i.e., who – themselves or the experimenter – had picked up the picture and named it). That is, the participants knew that X event had occurred, but they had difficulty in *remembering* the spatio-temporal context in which X occurred. Therefore, the metacognitive status of the personal memories reported by individuals with autism might justifiably be questioned, and cannot be taken as support for models 1 and 3.

A case of impaired metacognition but intact mindreading? Distinguishing model 2 from model 4. The autism research cited earlier supports equally models 2 and 4. As far as we can tell, once Carruthers concedes (sect. 3) that the mindreading system has different information available to it for the cases of self and others, respectively, the only way in which model 2 differs *theoretically* from model 4 is with respect to the role of

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

introspection of own propositional attitudes. Other commentators will no doubt debate whether it is possible to introspect our propositional attitudes directly. Here we concentrate on Carruthers' suggestion that different *predictions* emerge from models 2 and 4.

Williams and Happé (in press b; see also Williams 2008) addressed a potential confound within the classic "Smarties" unexpected contents task (Hogrefe et al. 1986), the task used most widely to assess awareness of false beliefs in self and others. In the Smarties task, participants *state* what they (falsely) believe is inside a Smarties box *before* they are asked the critical false-belief test questions. As such, it may be possible to answer the Self test question ('What did you think was inside the box, before you looked?') simply by remembering what one *said* and not necessarily what one *believed*. Although parallel performance across the Self and Other test questions of the task is usually observed among children with autism (e.g., Fisher et al. 2005), this potential confound may have led to an over-estimation of the ability of children with autism to reflect on their own false beliefs.

To test this possibility, we rigged a situation in which participants were asked by the experimenter (who feigned mild injury) to "get me a plaster," and had ready access to three different boxes: a plasters box, a crisps tube, and a sweets box. By opening the plasters box (which actually contained candles), participants demonstrated their (false) belief that the box contained plasters. However, having never verbalised their belief, success of the Self test question of this task ("What did you think was in the box, before you looked?") must reflect participants' recall of their false belief rather than of any prior statement.

We found that participants with autism were unique in finding the Self test question significantly harder than the Other-person question ("What will *x* think is inside the box, before s/he looks inside?"). Almost a quarter (21%) of our sample of 52 children with autism *failed* the Self question but *passed* the Other-person question. In contrast, less than 4% of participants with autism showed the opposite pattern of performance. Parallel performance across the test questions was observed in age- and ability-matched comparison participants, and in typically developing 3- to 5-year-olds. These results seem to show the kind of dissociation between mindreading and metacognition that Carruthers suggests would follow from model 2, but not from model 4.

In conclusion, we believe the data so far from autism support model 2 over model 4. But we are grateful for Carruthers' bold and exciting analysis, which helps to shape a new research agenda to answer the fascinating question: How well can people with autism "read their own minds"?

Making a case for introspection

doi:10.1017/S0140525X0900082X

Alexandra Zinck,[a] Sanne Lodahl,[b] and Chris D. Frith[b,c]

[a]*LWL-Universitätsklinik Bochum der Ruhr-Universität Bochum, Psychiatrie–Psychotherapie–Psychosomatik–Präventivmedizin, Institut für Philosophie, Ruhr-Universität Bochum, 44791 Bochum, Germany;* [b]*Centre of Functionally Integrative Neuroscience (CFIN), Danish National Research Foundation, and Institute of Philosophy and History of Ideas (IFI), and Faculty of Humanities, Aarhus University, Aarhus University Hospital, 8000 Aarhus, Denmark;* [c]*Wellcome Trust Centre for Neuroimaging, University College London, London, WC1N 3BG, United Kingdom.*

alexandra.zinck@rub.de

http://www.ruhr-uni-bochum.de/philosophy/staff/zinck/index.html

sanne@pet.auh.dk www.cfin.au.dk/menu478-en

cfrith@fil.ion.ucl.ac.uk http://www.fil.ion.ucl.ac.uk/Frith/

Abstract: Defending first-person introspective access to own mental states, we argue against Carruthers' claim of mindreading being prior to meta-cognition and for a fundamental difference between how we understand our own and others' mental states. We conclude that a model based on one mechanism but involving two different kinds of access for self and other is sufficient and more consistent with the evidence.

Making a case for introspection. Comparing four different accounts of the relationship between third-person mindreading (meta-representing mental states of others) and first-person metacognition (meta-representing one's own mental states), Carruthers concludes that the capacity to *mindread* is *prior* to metacognition. According to him, basic mindreading is either turned upon others or turned upon ourselves, the latter constituting metacognition. Mindreading is thus the capacity to interpret the other or the self and therefore does not require introspection.

This brings us to the core problem of our critique: Assuming that there is one basic meta-representational mechanism that underlies both understanding the self and other, how can this mechanism be characterized?

In what follows, our analysis hinges on the way Carruthers uses the term *introspection* in relation to basic *mindreading*. Most accounts of mindreading use introspection to describe a special kind of access that we have to ourselves that is not available for third-person mindreading.

Carruthers' account dispenses with this difference of access and the function of introspection. He gives a negative definition of introspection as "any reliable method for forming beliefs about one's own mental states that *is not* self-interpretative and that differs in *kind* from the ways in which we form beliefs about the mental states of other people" (sect. 1.4, para. 3, emphasis in original). Yet, in his architecture of the mind, there is no place for an introspective capacity constituting an immediate and direct inner perception of a belief. This conclusion results from Carruthers' extreme caution about the phenomenology that characterizes introspection and his dismissal of it as misleading.

This thesis of the unreliability of introspection and the necessity to dismiss it as a mode of access to beliefs is supported by data from confabulation and commissurotomy. However, this does not show that one cannot know one's beliefs to be true. We don't necessarily have to concede that beliefs are – or can become – consciously *uninterpreted*; instead we can assume that there are unconscious belief attitudes that can give rise to a conscious event whose content is a belief. In cases of confabulation, this doesn't mean, however, that we are not introspecting this event; it simply means that there is a discrepancy with the underlying belief attitudes.

So, the data suggests that there are mental processes that we are not conscious of. This is not a principled argument against introspective access to our own mental states that is independent of lengthy interpretation. The scope and quality of self-knowledge is limited, whether it is gained by introspection or by self-interpretation. Commissurotomy patients mistake their beliefs for certain actions, yet they do so also under an account of self-interpretation. Self-knowledge and its acquisition by introspection has certainly been overrated in philosophy, but the limitations are equal for self-interpretation.

Aside from this – although he claims that according to his account of mindreading applied to the self, there is "no . . . awareness of one's own propositional attitudes independently of any perceptually accessible cues that provide a basis for self-interpretation" (sect. 1.4, para. 2) – Carruthers does not completely differentiate introspection and self-interpretation according to his mindreading account and does concede there sometimes seem to be introspective qualities during self-interpretation, such as immediacy and effortlessness (sect. 8, para. 3).

So the question still remains of how best to characterize the access we have to ourselves. Contrary to Carruthers, we would like to argue that the immediacy and directness that characterizes

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

introspection is also present when mindreading others and that this is not a conscious interpretational endeavour. Just as with the perception of the outside world, our brain makes "unconscious inferences" (von Helmholtz 1866) when perceiving ourselves. This is the basis for the experience that introspection is direct and immediate. The same immediacy also occurs when perceiving others (Frith 2007). Nevertheless, there is a difference between the way we meta-represent our own and the mental states of others. When thinking of ourselves, there are more data available, that is, visceral and somaesthetic sensations in addition to a richer knowledge of our own past history. Thus, we are dealing with the same mechanism but with two different modes, one for the self and one for the other. This corresponds to Carruthers' model 2 account.

It accordingly also does not matter whether the mechanism evolved first for understanding others or for understanding oneself. We assume both involve the same underlying mechanism of meta-representation that, endowed with additional sources of information, makes up the different modes of access.

Another point of criticism against a *mindreading is prior* account is that the mechanism of mindreading is third-person directed. Thus, when I direct my mindreading capacity upon myself, I should use a third-person stance. Apart from being an interpretative process, this is also an unnecessarily complex and computationally expensive way of accessing the self. It can be argued that the best explanation is to simply accept the immediate first-person data instead of adopting the complex third-person setup.

A further argument for introspection as a specific mode of access for the self comes from considering why it might be valuable for survival: (1) we can inform others about our reasons for acting in a certain way; (2) we can gain high-level control of our emotion and our behaviour (e.g., Zelazo 2004). Take, for example, a simple learning process. We can learn associations between stimuli even when the stimuli are presented subliminally (i.e., not available to introspection). However, this learning is slow and gradual. If the stimuli are supraliminal, then insightful learning becomes possible through introspection. At some point subjects notice the contingency and will immediately jump to 100% performance (Pessiglione et al., in press).

Contrary to Carruthers, we prefer his model 2 that makes use of one mechanism but involves two different kinds of access: one which is perception-based for interpreting others, and additional introspective access which is available when assessing one's own mental states. Altogether, model 2 is more consistent and parsimonious. It also makes better predictions for pathologies such as autism and schizophrenia in which both kinds of access are impaired.

In sum, this discussion exemplifies that the understanding of how self and other are related is an important topic for research that is generating exciting new empirical and theoretical investigations.

Author's Response

Mindreading underlies metacognition

doi:10.1017/S0140525X09000831

Peter Carruthers

Department of Philosophy, University of Maryland, College Park, MD 20742.
pcarruth@umd.edu
www.philosophy.umd.edu/Faculty/pcarruthers/

Abstract: This response defends the view that human metacognition results from us turning our mindreading capacities upon ourselves, and that our access to our own propositional attitudes is through interpretation rather than introspection. Relevant evidence is considered, including that deriving from studies of childhood development and other animal species. Also discussed are data suggesting dissociations between metacognitive and mindreading capacities, especially in autism and schizophrenia.

R1. Introduction

The target article set out to consider four different accounts of the relationship between our mindreading and metacognitive abilities ("two independent mechanisms," "one mechanism, two modes of access," "metacognition is prior," and "mindreading is prior"). It argued in support of the fourth ("mindreading is prior") account, according to which metacognitive competence results from us turning our mindreading abilities upon ourselves. The target article considered a wide array of evidence bearing on the choice between the four accounts. This included evidence from childhood development, evidence from the role that metacognitive beliefs play in guiding human cognitive processes and behavior, evidence of confabulation in reports of one's own attitudes, alleged evidence of direct introspection of attitudes, comparative evidence of metacognitive competence in other species, evidence from autism, and evidence from schizophrenia. The commentaries take up an equally disparate set of topics. Some raise fundamental challenges that need to be confronted at some length, whereas others (as might be expected) are based upon misunderstandings. Table R1

Table R1.

No.	Section	Commentaries
R2.	The nature of the mindreading faculty	Buckner et al.; Friedman & Petrashek; Lurz; Pereplyotchik
R3.	The development of mindreading	Anderson & Perlis; Buckner et al.; Hernik et al.; Lewis & Carpendale; Rochat
R4.	The question of developmental priority	Fernyhough; Mills & Danovitch; Robbins
R5.	What is introspection?	Baars; Murphy; Pereplyotchik; Petty & Briñol; Zinck et al.
R6.	Evidence for and against introspection	Fiala & Nichols; Hurlburt; Zinck et al.
R7.	What is metacognition?	Anderson & Perlis; Couchman et al.; Kornell et al.; Proust
R8.	Metacognition in animals?	Couchman et al.; Kornell et al.; Proust
R9.	Dual processes and judgment	Buckner et al.; Frankish
R10.	The evidence from autism	Lombardo et al.; Williams et al.
R11.	Neuroimaging evidence	Lombardo et al.
R12.	The evidence from schizophrenia	Robbins; Wiffen & David
R13.	Some friendly suggestions	Evans; Huebner & Dennett; Langland-Hassan
R14.	Behaviorism bites back	Catania; Lewis & Carpendale
R15.	Conclusion	

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

provides a guide to the structure of the present reply, together with an indication of which commentaries are responded to (in whole or in part) in each section (including the notes attached to that section).

R2. The nature of the mindreading faculty

In the target article I had hoped that my argument in support of a "mindreading is prior" account of self-knowledge was independent of any specific commitments concerning the character of the mindreading faculty itself, beyond rejection of a simulation-based "metacognition is prior" alternative. I still think that is partly correct. Certainly I can accept that the mindreading faculty is not monolithic, but is actually a cluster of more specialized mechanisms working together in concert, somewhat as **Buckner, Shriver, Crowley, & Allen** (**Buckner et al.**) suggest. Indeed, this is what I actually believe, following Baron-Cohen (1995) and Nichols and Stich (2003). But one powerful objection to the proposal that there is no such thing as introspection for attitudes, raised independently by **Friedman & Petrashek** and by **Lurz**, has made me realize that the argument cannot be free of all such commitments. I shall first outline the objection, before showing how an independently motivated account of the architecture and mode of operation of the mindreading faculty can accommodate it.

The objection is that the mindreading system needs to have access to the agent's own beliefs in order to do its interpretative work; therefore self-attributing beliefs should be just as trivially easy as self-attributing experiences. **Friedman & Petrashek** claim, for example, that in order to form the metarepresentational belief that Bill believes that the first-aid box contains bandages, the mindreading system must access the attributor's own belief that first-aid boxes normally contain bandages. And they go on to stress that the mindreading system's default is to attribute the subject's own beliefs to other people, saying that this requires it to have access to those beliefs. Likewise, **Lurz** imagines a mindreader who observes a conspecific seeing some food. In order to draw any inferences from that fact, Lurz tells us, the mindreading system would have to access such beliefs as that the conspecific has recently eaten, or has shared food with others in the past. And again the moral is that the mindreading system needs to have access to the agent's own beliefs in order to do its work.

In light of these plausible claims, what might motivate one to deny that the mindreading system can access all of the agent's own beliefs? The answer is that the objectors forget about the frame problem. The idea that any single mental faculty might be conducting searches among all of a subject's beliefs is extremely problematic. Rather, there are likely to be a whole swarm of different decision-making systems that can conduct local searches of aspects of memory (Carruthers 2006). And a large part of the point of organizing cognition around a global workspace is that queries posted in that space can co-opt the resources of all the different consumer systems in parallel (Shanahan & Baars 2005). If the mindreading system is one of the consumer systems for global broadcast, as the target article assumes, then what we should predict is that it only has access to a limited database of domain-specific beliefs necessary to perform its computations.[1] But if this is so, then the challenge is to explain the datum that any one of one's beliefs can seemingly get appealed to in the course of mindreading.

To meet this challenge, I need to make two closely related distinctions. One is between System 1 mindreading (which is comparatively fast and done "online") and System 2 mindreading (which is slower, more reflective, and often involves supposition and simulation). This first distinction should need no defense. For reasoning about the minds of other people, like every other domain of reasoning that we know of, should admit of both System 1 and System 2 varieties. The other distinction is between verbally mediated forms of mindreading (such as answering a question about what someone believes) and kinds of mindreading that don't involve access to linguistic representations. (We can be quite sure that the latter exist, since even severely agrammatic aphasic people can retain their competence in nonverbal mindreading tasks. See Siegal & Varley 2002; Varley 1998.)

Consider, first, the fact that people will by default attribute their own beliefs to other people if asked. There is no reason to think that this requires the mindreading faculty to access those beliefs, any more than answering a question about one's *own* beliefs requires such access, as I argued in the target article (sect. 2.1). Rather, the executive and language-production systems cooperate (and partly compete) with one another, searching the attributor's own memory and issuing the result in the form of a metarepresentational verbal report – "I think/he thinks that P" – where the form of the report can be copied from the form of the initial question. The mindreading system has the power to intervene in this process when it possesses a representation of the target's belief that differs from the subject's own, but it plays no part in the default attribution process itself. Consistent with this suggestion, Apperly et al. (2007) show that people are significantly slower when responding to a probe about a target's false belief than they are when responding to a reality-probe.

Now consider a reflective, System 2, instance of mindreading (whether verbal or nonverbal). A query about the target's thoughts, goals, or likely behavior is posted in the global workspace (either in the form of a verbal question, or in an image of oneself in the situation of the target, say). The entire suite of consumer systems gets to work, drawing inferences and reasoning in their normal way, accessing whichever of the subject's beliefs they normally would, and the results are posted back into the global workspace once more, where they are accessible to the mindreading faculty as input, perhaps issuing in a conclusion or a further query. Here the entire process, collectively, has access to all of the agent's beliefs; but the mindreading system has access only to whatever gets posted in the global workspace (in addition to its own domain-specific database, of course, which is accessible to it when processing).

Finally, consider a case of "on-line" unreflective System 1 mindreading, of the sort that might be engaged in by the infants in the false-belief studies conducted by Onishi and Baillargeon (2005), Southgate et al. (2007), or Surian et al. (2007). Perceptions of the main aspects of the unfolding events are attended to and globally broadcast, thereby being made available to the full range of conceptual

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

systems including mindreading. These systems conceptualize and draw inferences from the input, with the former being fed back and broadcast as part of the perceptual state itself, and with the results of the latter being held briefly in the relevant domain-specific working memory system. (All System 1 reasoning systems will need to possess their own form of working memory, of course, to hold the results of previous computations while the next steps are undertaken. See Carruthers 2006.) Included in these broadcasts, then, will be the information that the target subject *sees* an object in one box rather than another, for example. And the working memory system that is internal to the mindreading faculty will contain such information as that the target *expects* the object to be where it was last seen and is *ignorant* of the fact that it has been moved. When combined with a novel perceptual input (e.g., the target subject returns on the scene after a brief absence), these beliefs enable an expectation to be generated concerning the target's likely behavior.

Notice that on this account no beliefs need to be accessible to the mindreading system beyond those residing in its domain-specific database, with the exception of those that are made perceptually available to it, on the one hand, and those that are immediately past products of its own operations, on the other. This is consistent with the fact that adults as well as children fail to take account of the mental states of other people in their online reasoning once the relevant facts are no longer perceptually salient and sufficient time has elapsed for any record to have been expunged from the mindreading system's working memory. See Keysar et al. (2003) for a dramatic demonstration of this point.

R3. The development of mindreading

Just as I had hoped to make the argument of the target article largely independent of assumptions about the nature of the mindreading faculty, so I had hoped to minimize assumptions about the latter's development. (Of course I do need to assume that development does *not* begin with first-person awareness of our own attitudes, in the way that Goldman [2006] suggests.) In particular, I tried to steer clear of the dispute between nativist or "core knowledge" approaches, on the one hand (e.g., Fodor 1992; Leslie et al. 2004) and constructivist or theorizing-theory accounts, on the other (e.g., Gopnik & Melzoff 1997; Wellman 1990). Here, too, my efforts were partly, but by no means entirely, successful, as I now explain.

Both **Hernik, Fearon, & Fonagy** (**Hernik et al.**) and **Rochat** emphasize the crucial roles played by emotion and emotional engagement with others in the development of mindreading; and each appears to think that this claim conflicts with some aspect of the target article. But I see no problem with accepting these data. Both nativists and theorizing-theorists can believe in the developmental importance of emotional engagement, but will interpret the sources of that importance differently. Moreover, no evidence is provided that an understanding of one's own emotions precedes an understanding of the emotions of others in development (which would be problematic for a "mindreading is prior" account). On the contrary, Rochat writes: "From a developmental vantage point, affective reading and meta-affectivity are ontologically linked, representing two sides of the same coin." This is, of course, further grist for my mill.

Construed as a thesis about the architecture of the mature mind, the "mindreading is prior" account is independent of the debate between nativists and theorizing-theorists about the development of the mindreading system. But **Buckner et al.** are correct in pointing out that one (though only one) of the *arguments* that I use in support of the "mindreading is prior" architecture depends upon some or other variety of nativist position (whether this be an innately given body of knowledge, or an innate module, or an innate domain-specific learning mechanism). For I claim that there is a good evolutionary explanation of the emergence of mindreading in highly social creatures such as ourselves, whereas there are no good evolutionary reasons for the emergence of introspection for attitudes (or else those reasons makes predictions that are not borne out by the metacognitive data, either human or comparative). This is supposed to count in favor of a "mindreading is prior" account. And it plainly commits me to some or other version of nativism about the course of mindreading development.

Buckner et al. argue, in contrast, that metarepresentational mindreading may be a late exaptation of more primitive capacities, grounded in these together with our linguistic abilities and general-purpose concept-learning and theorizing skills. They think that the only true adaptations in the social-cognitive domain are a swarm of first-order, non-metarepresentational, mechanisms for face recognition, eye-tracking, automated imitation via the mirror neuron system, and so forth. But there are two main problems with this view. One is the rapidly expanding body of evidence of *very* early metarepresentational competence in infants, embracing false-belief understanding inter alia (Bosco et al. 2006; Onishi & Baillargeon 2005; Onishi et al. 2007; Song & Baillargeon, forthcoming; Song et al., forthcoming; Southgate et al. 2007; Surian et al. 2007). And not all of these studies, it should be stressed, use looking time as a measure of expectation violation. On the contrary, Southgate et al. (2007) use anticipatory looking as their dependent measure, which is much less ambiguous.

The other major problem with **Buckner et al.**'s suggestion is that mindreading is required in order to learn a language in the first place. I don't deny that syntax may be innate, or acquired through the offices of a dedicated domain-specific learning mechanism. But learning the lexicon requires children to figure out the referential intentions of the speakers around them (Bloom 2002). And this plainly requires metarepresentation. Moreover (and just as this account predicts), we have ample evidence that infants can attribute goals and intentions to others in the first year of life, significantly before they can attribute beliefs and misleading appearances (Csibra et al. 2003; Johnson 2000; Luo & Baillargeon 2005; Woodward 1998).

Buckner et al. write admiringly of the work of Gallagher (2001; 2004) in this connection, as do **Anderson & Perlis** and **Lewis & Carpendale**. But Gallagher's work is subject to both of the objections just outlined. Moreover, he goes awry in his critique of the opposing approach, to which he refers with the generic "theory-theory" (intended to cover both nativist and theorizing-theory varieties). In particular, it is simply false that theory-theorists must (or do) assume that mentalizing usually involves the

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

adoption of a third-person, detached and observational, perspective on other people. On the contrary, theory-theorists have always emphasized that the primary use of mindreading is in *interaction* with others (which Gallagher calls "second-personal"). That, after all, is what "Machiavellian intelligence" is all about. And the fact that our apprehension of the meaning of other people's behavior is often phenomenologically immediate does not, of course, show that it isn't underpinned by theory-driven computations of underlying mental states. Indeed, there is simply no other way of explaining our competence in this domain. Appealing just to sensory-motor skills (as Gallagher does) is plainly inadequate to account for the *flexibility* of the ways in which adults and infants can interact with others. Indeed, in order to interact flexibly with *any* complex system (be it physical or human), you need a good enough understanding of how it works.

R4. The question of developmental priority

The target article expressed skepticism about the capacity of developmental data to discriminate between a "mind-reading is prior" account and its three main competitors (sect. 4). **Mills & Danovitch** disagree. They cite a number of forms of evidence suggesting that mindreading skills of various sorts emerge in development prior to metacognition, which they say supports a "mindreading is prior" account. Since one of my main grounds for skepticism concerned arguments for the priority of mindreading that are premised on the *parallel* emergence of mindreading and metacognition in development (which fails to discriminate between the "mindreading is prior" view and the "one mechanism, two modes of access" account), I am happy to agree. But let me sound the following cautionary note. Until we have a good understanding of the reasons for the two-year developmental lag between children's capacities to pass nonverbal and verbal versions of mindreading tasks, arguments that rely upon the latter need to be treated with some caution. For it may be that the "self" and "other" versions of a verbal task differ along whatever turns out to be the relevant parameter. Put differently: you can't control for confounding factors that you don't yet know about.

In response to **Mills & Danovitch**, I should also stress that although a finding that mindreading competence is developmentally prior to metacognition would support a "mindreading is prior" account (because it would be inconsistent with the other three alternatives), this is not actually a *prediction* of the account. For the latter claims only that it is *the same system* that underlies our mindreading capacity that gets turned upon ourselves to issue in metacognition. It does not claim that the first occurs in development before the latter. (In this respect, the label "mindreading is prior" may be somewhat misleading. I intend it only to refer to a *functional* and/or *evolutionary* priority.)

Fernyhough would plainly disagree with the point made in the previous paragraph. He gives reasons for thinking that it may take time for aspects of children's inner lives to develop. In particular, the transformation of private speech ("talking to oneself") into *inner* ("silent") speech may not be complete until middle childhood; and capacities to build and sustain visual images may likewise be slow to develop. Because the target article claims that these are among the data that the mindreading system uses when attributing propositional attitudes to oneself, Fernyhough says that the "mindreading is prior" account must therefore predict that metacognition should lag significantly behind mindreading in development. But there is no such implication. All that follows is that there will be many more moments in the daily lives of children at which they will be unwilling to attribute occurrent thoughts to themselves than is true of the daily lives of adults, because the conscious mental events that might underlie such self-attributions simply are not present. Nothing follows about children's *competence* to self-attribute attitudes. Nor does it follow that children will be weaker at attributing attitudes to themselves than they are at attributing attitudes to others, provided that the tasks are suitably matched.

Robbins claims that I have overlooked crucial conflicting evidence, which demonstrates that metacognition is prior to mindreading in development. He cites a study by Wimmer et al. (1998) which seems to show that young children have awareness of their own knowledge before they have awareness of the knowledge of other people. But the study in question admits of an alternative explanation. In the "self" condition, the children are allowed to look, or not look, into a box, and are then asked whether they know what is in the box; whereas in the "other" condition they observe a subject either looking, or not looking, into the box before being asked whether the subject knows what is in the box. Answering the question in the "other" condition requires the children to reason appropriately from the generalization that seeing leads to knowing (or some such). But answering the question in the "self" condition requires no such thing. The children can answer simply by accessing, or by failing to access, their knowledge of what is in the box. They can substitute a first-order question in place of the second-order question asked – namely, "What *is* in the box?" – and answer "Yes," that they do know what is in the box, if an answer comes to mind, otherwise answering "No."

R5. What is introspection?

Baars thinks that the target article is committed to denying that nonhuman animals and young infants feel pain. This is because these subjects are incapable of mindreading, and because the target article denies the existence of introspection. But there are two distinct misunderstandings at work here.

One is based upon an unfortunate ambiguity in the use of the term "introspection." In one sense, introspection is any form of looking within *the body*. In this sense, perceptions of pain or of one's own beating heart count as introspections. In another sense, introspection is a form of looking within *the mind*. In this sense, the outputs of an introspective process are always metarepresentational, involving representations of one's mental states as such. And in this sense, perceptions of pain or of heartbeat are definitely *not* introspections, since they issue in first-order representations of properties of the body. It should be stressed that it is only this latter, metarepresentational, sense of "introspection" that is at stake in the

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

target article. Hence, even if I denied the existence of introspection in this sense altogether, there is no reason why this should commit me to denying that animals feel pain, or fear, or hunger, or thirst. For what is in question in these cases is only introspection in the first "within the body" sense.

Baars's second misunderstanding lies in believing that the target article denies the existence of introspection (in the metacognitive sense) for all categories of mental state. For he thinks that the view will have difficulty in accounting for the reliability of metacognitive self-report in psychophysics. But I specifically allow (indeed, I insist) that globally broadcast perceptual and quasi-perceptual states can be introspected, because they are available as input to the mindreading faculty. Self-attribution of such states should therefore be trivial for anyone who possesses the requisite concepts, which can then be applied to the input-states on a recognitional (non-interpretative) basis.

Pereplyotchik, too, misunderstands the sense of "introspection" that is at issue. For he thinks that it will be sufficient to demonstrate that there is no such thing as introspection for perceptual states if it can be shown that the mindreading system relies upon a tacit theory in self-ascribing such states. This is a mistake. That a process is introspective is *not* supposed to be inconsistent with it involving computations or inferences of various sorts (provided they are unconscious ones), so long as the inferences rely only on information of a general kind, and do not access information about the agent's circumstances, behavior, or earlier mental states. For remember, what is at stake is whether our access to our own minds is different *in kind* from our access to the minds of other people. And the latter always involves just such inferences. This was also the reason why I defined introspection *negatively* for the purposes of the target article. For I wanted to leave open "inner sense" accounts as well as "application of a tacit theory" views of introspection, according to each of which the attribution of mental states to oneself is inferential (but still quite different from the attribution of mental states to other people).

Zinck, Lodahl, & Frith (**Zinck et al.**) mistake the nature of the intended contrast between a "one system, two modes of access" account and a "mindreading is prior" view. They insist that when metarepresenting our own mental states, the mindreading system has access to a richer array of data, such as visceral and somaesthetic sensations, and that this therefore supports a "one system, two modes of access" account. But I, too, claim that the mindreading system can utilize data when attributing states to the self that are not available when attributing the same states to others, and I maintain that this is consistent with a "mindreading is prior" view. As I intend the distinction, the difference between the two forms of account is *not* whether there are different *data* available to the mindreading system when attributing mental states to oneself or to another. Rather, the difference concerns whether the mindreading system employs two different *informational channels* in the two cases. The distinction is intended to be an architectural one.

Because the mindreading system utilizes the very same mechanism of "global broadcast" of attended outputs of perceptual systems, whether attributing mental states to oneself or to another, this means that there are *not* two different modes of access to mental states, even though the perceptual and quasi-perceptual states that are utilized in the two cases are often different. To provide evidence supporting a "one system, two modes of access" account, **Zinck et al.** would need to show that we can self-attribute propositional attitude states independently of any sensory or imagistic information accessible via global broadcast. But they themselves seem to doubt whether any such evidence exists.

Murphy denies that the mindreading system is always implicated in our knowledge of our own attitudes, while agreeing with me that there are no special mechanisms that enable us to detect and describe those attitudes. Rather, he thinks that we can do whatever we would normally do to determine a truth about the world and can then use the result of that same process to self-ascribe the resulting belief. This might well work as an account of how we express our beliefs in speech. Indeed, so construed, it is an account that I endorsed in the target article (sect. 2.1). The language production system can take as input the result of a belief-forming process, or the result of a search of memory, and can formulate that input-state into a belief report. We can imagine this happening via a two-step process: the language system accesses a belief with the content *P* and draws on lexical and syntactic resources to express this in a sentence, "P," before elaborating it and articulating the result in the form, "I believe that P." But I deny that such an account can succeed as an account of meta*cognition*, or as an account of how we form *beliefs* about our own beliefs.

Murphy is confronted with the following dilemma. Suppose, first, that the assertion, "I believe that P" is an encoding into language of a previously existing metacognitive belief (the belief, namely, that I believe that P). Then the challenge is to explain how this belief is arrived at without either implicating the mindreading faculty or appealing to any special introspective channel. But there would seem to be just two possible ways for whatever process that issues in such a belief to do its work (in a reliable enough way). One would be for it to have access to the output of the process that issues in the belief or memory that P (which would then surely involve some sort of introspective channel of access to the latter). The other would be for the metacognitive belief-forming process to involve interpretation and inference from other events, such as a prior tokening of the assertion, "P," or the occurrence of a memory-image caused by the belief in question (which would surely then implicate the mindreading faculty, or else some system with many of the same powers as the mindreading faculty).

So **Murphy** must intend (much more plausibly) that the assertion, "I believe that P" can be generated directly from the belief that P, without subjects first needing to form the metacognitive belief that they believe that P. As described above, the language system (working in concert with executive systems, no doubt) can access the belief that P but then formulate this into the sentence, "I believe that P," rather than the first-order sentence, "P." But this assertion is not *itself* a metacognitive belief (nor, by hypothesis, does it involve one). Rather, it is a linguistic action (albeit one with a metarepresentational content). The system that issues in the metacognitive belief that I believe that P must take this assertion as input and deliver the metacognitive belief as output. But in order to do this, it would have to

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

engage in interpretation, just as when hearing a similar assertion made by another person. Because the assertion could be a lie, or be meant ironically, or meant as a joke, it is hard to see how the necessary interpreting could be done except by the mindreading faculty (or else some system with many of the same powers as the mindreading faculty). But this is now the view endorsed by the target article: In such cases I come to know what I believe by hearing and interpreting what I say (whether overtly or in inner speech). Murphy has failed to present us with a genuine alternative.

Petty & Briñol agree with the target article that self-attributions of attitudes always involve interpretation. But they insist that interpretation is a matter of *degree*, and that sometimes interpretation can be so minimal as to be almost indistinguishable from introspection. I agree with the former point but not with the latter. Of course it is true, as Petty & Briñol point out, that there is a big difference between interpretations of oneself that rely only on publicly available information (such as one's own behavior and circumstances) and interpretations that rely only on subjectively accessible mental events (such as one's own somatic feelings and/or one's own "inner speech"). But the main point at issue in the target article *is* a dichotomous, architectural one. It concerns the existence (or not) of a distinct informational channel to our own attitudes, different from the sensory channels that are available to the mindreading system for use in interpreting other people. There either is such a channel or there is not. (The target article claims the latter, and Petty & Briñol appear to agree.) Moreover, even minimal-interpretation cases are much less similar to introspection than Petty & Briñol seem to think. Consider their example of someone who says to himself, "It is good," when tasting some ice-cream, and thereby interprets himself as liking ice-cream. The mindreading faculty, functioning together with the language comprehension system, has to fix on the object of evaluation ("*What* is good?"), interpret the evaluative predicate ("In what sense is it good?"), and determine what sort of speech act is being expressed (whether literal, suppositional, ironic, or whatever). No doubt the answers can, in context, be settled quite easily. But they are exactly *the same* answers that would need to be provided when interpreting the speech of another person. And no one should think that the latter is at all similar in its nature to introspection.

R6. Evidence for and against introspection

Fiala & Nichols challenge the claim made in the target article that confabulators often have the impression that they are introspecting rather than self-interpreting (sect. 3.1), which is a crucial component of the argument against introspection for attitudes. They first point out that no one has ever *asked* a split-brain subject whether or not he thinks he is introspecting. But this would be a bad question to ask, for a number of reasons. One is that "introspection" is a term of art, and requiring people to make judgments involving an unfamiliar term is unlikely to be a reliable way of finding out what they believe. Another is that the direct-question method is a poor way of accessing people's tacit beliefs in general (Scholl 2007). I doubt that many people have explicit, verbalizable, beliefs about the nature of their access to their own mental states – with the possible exception of those who have taken an introductory course in philosophy. Rather, the way in which people think and reason about their own mental states just *assumes* that the latter are transparently accessible to them. But if *asked* about that access, who knows what they might say? For they will almost certainly find the question to be confusing, and they might revert to bits and pieces of knowledge acquired about Freud, or about cognitive science, or whatever, when trying to say something sensible by way of answer.

So what is really in question is whether it seems to split-brain subjects that they are formulating beliefs about their own mental states and processes in whatever way they normally would – in a way that doesn't seem to them to involve self-interpretation – not whether they have explicit beliefs about the process in question. This is hard to assess directly. But those who work with such people say that their own sense of themselves following the split-brain operation seems to be unchanged (Gazzaniga 1995). And even reminders of their split-brain status that are made immediately prior to testing – and that are given, moreover, to those who have a good theoretical understanding of the effects of the operation – have no effect (Gazzaniga, e-mail communication, November 8, 2006). The subject goes right on confabulating. This isn't what one would predict if subjects were, at any level, aware of interpreting themselves, since one would expect that a reminder of their split-brain status should enrich their hypothesis pool. But it does not.

Fiala & Nichols also point out that there are many examples from the confabulation literature where subjects express their metacognitive thoughts with low confidence, suggesting that they are not only interpreting themselves but are at some level aware that they are doing so. The point is entirely correct. But it doesn't have the consequences destructive of my argument that Fiala & Nichols allege. This is because there are also a great many instances in which subjects express their metacognitive beliefs unhesitatingly and with *high* confidence. And these are all that I require to make my case. Indeed, the self-interpretative model of attitude self-awareness *predicts* that there should be cases of both sorts. For only if an interpretation can be arrived at smoothly and unhesitatingly will subjects have the impression that they are introspecting. In more problematic cases such as those that Fiala and Nichols cite, or such as especially bizarre actions performed following hypnosis, it will be more difficult for the mindreading system to generate an interpretation (just as it would be difficult to interpret such behavior observed in another). And as soon as subjects become aware of themselves as interpreting, they are likely to express any belief that they formulate with some caution.

Note that exactly the same distinction can be made with respect to other-person mindreading. In many cases the interpretation process is swift and unconscious, and the resulting phenomenology is that we just seem to *see* someone's behavior as informed by certain beliefs and goals. (Here I am in full agreement with **Zinck et al.**) But in other cases an interpretation is harder to come by, and we become aware that we are trying to interpret. (See also the discussion of System 1 versus System 2 mindreading in sect. R2.)

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

In the target article I assumed that one of the biggest challenges to a "mindreading is prior" account derives from the "descriptive experience sampling" studies conducted over the years by Hurlburt and colleagues (Hurlburt 1990; 1993; Hurlburt & Akhter 2008; Hurlburt & Heavey 2006), specifically the finding that subjects will sometimes report engaging in "unsymbolized thinking" at the time of the beep. I took this to be evidence that subjects are capable of introspecting their propositional attitudes, and tried to respond. However, **Hurlburt** now replies that I have misinterpreted his position. Unsymbolized thoughts are merely thoughts that don't have any semantically relevant images, words, or other sensations as the "primary theme or focus" of the subject's attention at the time of the beep. Hurlburt concedes that such experiences are generally present in the periphery of attention, providing a basis for self-interpretation. Moreover, he argues that the ways in which subjects respond when probed about these episodes actually speaks *in favor* of a "mindreading is prior" position on our awareness of our own attitudes. This additional support from an unexpected quarter is, of course, most welcome.

R7. What is metacognition?

A number of commentators accuse me of using the term "metacognition" in a non-standard sense (**Anderson & Perlis**; **Couchman, Coutinho, Beran, & Smith** [**Couchman et al.**]; **Proust**).[2] These commentators allege that the normal usage in cognitive science is that metacognition is involved in any process that has a controlling influence on the way that another cognitive process unfolds. On this account, it is left open whether or not metacognition need involve metarepresentations of the events within the cognitive process that gets controlled.

I am happy to allow that some authors might use the term in this (hereafter "control") sense. But I deny that it is a common – let alone a standard – usage. In general in the metacognition literature in psychology, metacognition is defined in terms of thought about our own thoughts. Indeed, **Proust** herself provides the standard definition (Proust 2007, p. 271): "This is the domain of metacognition: thinking about one's own thinking." (See also Dunlosky & Metcalfe 2009; Flavell 1979; Koriat 2007.) And it is then a matter of substantive *investigation* whether or not, and to what extent, metacognition has a controlling function. (See especially Koriat et al. 2006.) This wouldn't even make sense if metacognition were *defined* in terms of control.

It is important to emphasize that the control and metarepresentational senses of "metacognition" are two-way independent of one another. There are certainly many instances in which one cognitive process exercises a causal influence on another without the former involving any metarepresentations of any aspect of the latter. (See sect. 5.1 of the target article for some examples.) And in connection with any metarepresentational form of metacognition, it will always be an open question whether or not it has a causal influence upon the cognitive state or process represented. Although these points are very well understood by most researchers, some are apt to think that they can move freely from talk of metacognition in the control sense to metacognition in the metarepresentational sense. This is especially true of some of those who work in the field of animal metacognition. Some, I think, are quite clear-headed that they are seeking forms of metacognitive control for which the *best available explanation* will be the occurrence of a metarepresentational process. (See especially Metcalfe 2008; Son & Kornell 2005.) But some seem unaware that any additional argumentation is needed to get from metacognition in the control sense to metacognition in the metarepresentational sense. This is especially true of the commentary by **Couchman et al.**, as well as the articles by members of their team cited therein, which I discuss in section R8.

Proust raises a more substantive challenge to the assumptions of the target article. She suggests that the latter overlooks the possibility of *nonconceptual* forms of metacognition (in the metarepresentational sense of the latter term). Specifically, she suggests that epistemic feelings like surprise and confidence should be seen as nonconceptual representations of the underlying mental states (such as violated expectations or high degrees of belief). Hence, any person or animal that can use such feelings as a *cue* to guide further behavior (such as looking more closely at the target event) can be said to be acting as a result of a metacognitive process. This is an interesting idea, which deserves examination. It will require us to delve a bit into competing theories of the nature of intentional, or representational, content.

Let us assume (with **Proust**) that epistemic feelings like surprise and confidence are distinctive forms of somatosensory experience that are caused by an underlying cognitive state or process, but without involving any conceptualization of that state or process as such. So an animal that feels surprise has an expectation (a belief) that is violated by what it is currently perceiving, which in turn causes a suite of bodily reactions of which the animal is aware (heightened alertness, widening of the eyes, automatic orienting towards the stimulus, and so on), but without the animal necessarily knowing *that* it has an expectation that has been violated. Because the epistemic feeling is reliably caused by a cognitive state or event, it thereby carries information about it. And then on any purely informational account of representational content (e.g., Fodor 1990), the feeling can count as a nonconceptual representation of the representational state or event in question (that is, it counts as a metarepresentation). One problem with this proposal, however, is that it makes metarepresentations come too cheap. For almost all mental states, processes, and behaviors will carry information about the existence of some other mental state or process, thereby becoming nonconceptual metarepresentations of the latter, on the proposed account. Thus inferential processes will characteristically carry information about (and hence metarepresent) the presence of beliefs, decision-making processes will carry information about the presence of beliefs and desires, and so forth.

Moreover, few researchers in cognitive science actually rely upon an informational account of representation in their own work. Most adopt some or other variety of inferential or conceptual role semantics (e.g., Block 1986), according to which what a symbol represents depends (at least partly) upon the use that the rest of the cognitive system is apt to make of that symbol. This is probably wise, because purely informational accounts of intentional

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

content face notorious difficulties (one of which will be mentioned further on; see Botterill & Carruthers 1999 for discussion). And then the question for us becomes: Does the animal *make use of* the epistemic feeling in question in such a way that the feeling is thereby constituted as a nonconceptual representation of a cognitive state?

Consider, first, paradigmatic cases of nonconceptual representation, such as a perceptual representation of a colored surface or of the detailed shape of an object. In virtue of what does the perceptual state represent the colored surface rather than, for instance, a particular pattern of activity on the retina or in the optic nerve (since it equally carries information about both)? A natural answer is that the animal itself *treats* that representation as a representation of color – it thinks and acts in the sort of way that would be appropriate if it *were* a representation of the color of a surface. For example, perceiving the red surface of a fruit, and believing that red fruits of that type are ripe, the animal might grasp and eat it. Likewise, a perceptual representation of the detailed shape of an object will be used to guide the animal's choice of grip size and hand orientation when it reaches out for it. It seems that a nonconceptual representation of some property of the world represents what it does partly in virtue of its role in guiding thought and action that is focused on that aspect of the world.

Consider, now, epistemic feelings, such as the feeling of low confidence that an animal might experience when faced with an especially difficult judgment or discrimination. This is a feeling that involves an aversive state of anxiety, caused by the animal's low degree of belief. Should it be considered a nonconceptual representation of a cognitive state (one of low degrees of belief or of conflicts of belief), as **Proust** suggests? To answer, we need to look at how it is used by the animal. One thing that the animal might do in consequence is opt for a high-confidence, low-anxiety, option instead. But this is an action that is targeted *on the world* rather than on the animal's own beliefs. It should lead us to say that the feeling of uncertainty is a representation of the riskiness of certain worldly options or events, rather than a representation of the animal's own low degree of belief. For the animal doesn't act in a way that is directed at its own beliefs; rather it acts on the world. Likewise for an animal that is led by its feeling of uncertainty to engage in information-seeking behavior such as examining the object more closely, walking around it to look at it from the other side, sniffing it, pressing a "hint" key of the sort employed by Kornell et al. (2007), and so on: These are behaviors that are aimed at answering a first-order question about the object – "Is it edible?", "Is it safe?", "What comes next?", and so on – rather than being aimed at changing the animal's own degrees of belief. It seems reasonable to conclude, therefore, that epistemic feelings should not be regarded as possessing metarepresentational nonconceptual content.[3]

Moreover, there is no reason to think that epistemic feelings are a first evolutionary step on the road to metarepresentation. This is because metarepresentation requires the development of concept-wielding consumer systems for the bodily cues in question, which contain implicit or explicit theories of the nature and causal roles of the underlying mental states. (Note that even a simulation theorist like Goldman [2006] needs to postulate an innately structured set of representations in a language of thought linked up to the different mental state kinds.) It should be stressed that the bodily feelings in question – that are distinctive of surprise, or the anxiety that attends uncertainty, for examples – are just that: bodily feelings. By themselves they give no clues as to the nature of the mental states that cause them (a violated expectation, in the case of surprise, and low or conflicting degrees of belief, in the case of uncertainty). How would an animal that as yet had no conception of those types of mental state be expected to acquire one? Certainly not via individual learning. And if via evolution, then it is far from clear where the pressure to develop such theories is to come from. Not from the benefits of metacognition in the control sense, presumably, since by hypothesis the animals in question already have that (see sect. R8). Hence, the pressure is presumably social, in which case what develops will be a mindreading system (albeit one that is capable of taking bodily cues as input).

R8. Animal metacognition?

There are two distinct ways in which an animal might behave in metacognitive fashion (in the control sense) without engaging in metacognition (in the metarepresentational sense). First, it might utilize degrees of belief and desire (without metarepresenting them as such, of course), combined with one or two simple first-order (non-metarepresentational) mechanisms and/or acquired first-order beliefs. This is the explanatory strategy followed in Carruthers (2008b) and described briefly in the target article. But second, as **Proust** explains, an animal might utilize its own bodily changes and reactions (including feelings that are distinctive of surprise, uncertainty, and familiarity, for examples) as *cues*. Thus, an animal might be innately predisposed, or might have learned, that the bodily feeling distinctive of uncertainty is a cue to engage in actions that obtain improved information (e.g., by moving its head from side to side for a better view), or to engage in actions that have the effect of "opting out" of the current situation and entering a new one. (Actually, this might be one way of *implementing* the "gate-keeping" mechanism postulated in Carruthers 2008b, which takes roughly equivalently motivated but incompatible plans of action or inaction as input, and which motivates a search for additional information or alternative behavior.) Note that on this second alternative, the animal does not have to *recognize that* it is surprised or uncertain; indeed it need not even possess the concepts of surprise or uncertainty. Rather, it just has to be capable of recognizing a distinctive bodily gestalt or feeling, and initiating an innately prepared or learned response. (Nor, as sect. R7 argues in reply to Proust, does the feeling itself possess a nonconceptual metacognitive content.) Such an animal might display any of the metacognitive control-behaviors currently investigated by comparative psychologists, as I shall show briefly below. But it might be wholly incapable of metacognition in the metarepresentational sense.

Carruthers (2008b) describes exhaustively and in detail how all of the data cited in the commentary by **Kornell, Schwartz, & Son** (**Kornell et al.**) are explicable in non-metarepresentational terms, utilizing *degrees* of attitude

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

strength. Kornell et al. make no attempt in their commentary to respond to those arguments, or to demonstrate why their own metarepresentational interpretation of their data is superior. It wouldn't be fruitful just to repeat the arguments of my 2008b paper here. Rather, I shall show briefly how an animal that can treat its own epistemic feelings as a cue might behave in the ways that Kornell et al. describe without being capable of metarepresentation. Thereafter I shall show how **Couchman et al.** chronically conflate the control and metarepresentational senses of metacognition in their commentary and other work.

Consider what is perhaps the most challenging set of data presented by **Kornell et al.**, citing Kornell et al. (2007). Rhesus monkeys were initially trained in a gambling task that required them to first make a difficult perceptual discrimination before choosing between a "high-risk" and a "low-risk" symbol. Pressing the former would issue in a large reward if the initial discrimination was correct but a large loss if the discrimination was incorrect; pressing the latter would issue in a guaranteed small reward. The monkeys were then trained on a quite different – memory – task (initially without the gambling option). Following training, when the "high-risk" and "low-risk" symbols were re-introduced, the animals immediately made adaptive use of them. Kornell et al. take this to show that the monkeys had learned a general metacognitive strategy for solving the tasks. Carruthers (2008b) shows how these data can be better explained in terms of degrees of belief combined with a capacity for abstract rule-formation. Here let me sketch a further alternative: that the animals might have learned to use their own feelings of uncertainty as a cue.

We can presume that monkeys are capable of both being, and *feeling*, uncertain, even if they are incapable of metarepresentation of any sort. The monkeys in the first phase of the experiment just described could then have learned to treat their own feeling of uncertainty when making an initial discrimination as a cue to press the "low-risk" symbol thereafter. They would therefore have acquired, and learned to act upon, a rule of the form, "When *that* bodily feeling/gestalt is present, press the 'low-risk' symbol when it arrives." (Note that there is nothing metarepresentational contained here. The feeling in question is a state of the body, not of the mind. See sects. R5 and R7.) When the monkeys then entered the second phase of the experiment they would, of course, sometimes feel uncertain, but this time whenever they were presented with a difficult *memory* task. The introduction of the gambling option might then have activated, and led them to act upon, the exact same rule.

I now turn to consider **Couchman et al.** It is plain that at the outset of their commentary they actually use "first-order" to mean "behaviorist," and that by "metacognitive" they mean any process that is genuinely *cognitive*, with the animal taking decisions in light of its beliefs. For they describe Smith et al. (2006) as supporting a "metacognitive" account. In those experiments both feedback and rewards were deferred until the animal had completed a block of trials, thus preventing the creation of stimulus-response pairings that might otherwise explain the animals' adaptive use of the uncertainty response. Couchman et al. write, "It was clear in that study that monkeys' uncertainty-response strategies were adjudicated cognitively and decisionally, not using first-order cues." I agree (at least, if by "first-order cues" one means "stimulus-response pairings"). But the training would have given the animals ample opportunity to acquire a set of non-metarepresentational beliefs about the contingencies of the experiment. By the time that they entered the test phase, they would know that pressing the "dense" key if the stimulus was dense would thereafter issue in a reward, whereas pressing the "dense" key if the stimulus was sparse would thereafter issue in a penalty, and that pressing the "uncertain" key would issue in neither a reward nor a penalty. These beliefs, combined with *degrees* of belief that a given stimulus is dense, or sparse, can then explain the data in an entirely non-metarepresentational way, as Carruthers (2008b) demonstrates.

Couchman et al. point out, quite correctly, that the non-metarepresentational explanation adverted to in the foregoing requires the postulation of what Carruthers (2008b) calls a "gate-keeping mechanism" (which might be absent in capuchins and pigeons, note, thus accommodating the findings of Beran et al. [in press] and Inman & Shettleworth [1999] that neither species makes adaptive use of an uncertainty response). This is a mechanism that is sensitive to the presence of beliefs or motivations for action of roughly equal strength, issuing in a search for additional information or alternative strategies when receiving such states as input. Couchman et al. object that this commits me to a metacognitive explanation of the data, and they write, "It [the gatekeeper mechanism] meets the definition of a second-order, controlled cognitive process." Since it is plain that the mechanism in question need not involve any metarepresentations for it to operate as envisaged, Couchman et al. must here be using "metacognitive" in the control rather than the metarepresentational sense.

So far there isn't any substantive disagreement between **Couchman et al.** and myself, just "crossed wires" resulting from differences in the use of the term, "metacognitive." But they go on to conclude their commentary by claiming victory for a "metacognition is prior" account over my own "mindreading is prior" model, despite the fact that the two are perfectly consistent with one another if the former is taken in their control sense and the latter is understood in my metarepresentational sense. They also offer an account of the origins of mindreading that is blatantly and explicitly Cartesian, presupposing that we have prior awareness and understanding of our own mental states *as such* (i.e., presupposing the prior existence of metacognition in the metarepresentational sense). I fear that Couchman et al. have engaged in a fine body of experimental work that is framed and guided by theoretical confusion.

R9. Dual processes and judgment

Frankish takes issue with the argument of section 7 of the target article, which claims that the conscious events that take place at the System 2 level (e.g., verbalizing to myself, "P," or, "I shall do Q") don't have the right kind of causal role to constitute a judgment or a decision. For they only achieve their effects via further (unconscious) processes of reasoning. So although these events are introspectable, this doesn't mean that any judgments or

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

decisions are introspectable. Frankish replies that these events have a *System 2 role* appropriate for a judgment or decision. For they are the *last System 2 events* that occur prior to the characteristic effects of judgments and decisions. While he acknowledges that further reasoning processes of a System 1 sort occur subsequent to those events, mediating their causal effects on behavior, he says that these should be thought of as belonging to the *realizing base* of a System 2 judgment or decision.

However, our commonsense notions of *judgment* and *decision* don't make any allowance for the System 1/System 2 distinction. A judgment is a content-bearing event that gives rise to a stored belief with the same content *immediately*, and which is likewise immediately available to inform practical decision-making, without the intervention of any further reasoning. Similarly, a decision is a content-bearing event that causes intention or action without the mediation of any further reasoning about whether or not to act. By these lights, neither the judgment-like event of saying to myself, "P," nor the decision-like event of saying to myself, "I shall do Q," can qualify. Moreover, while it may be true enough that System 2 processes in general are realized in those of System 1 (Carruthers 2009), the realizing conditions for a particular event surely cannot occur subsequent to that event itself. And yet it is only once the conscious events of saying to myself, "P," or, "I shall do Q," are completed that the System 1 reasoning leading to belief or action kicks in. In addition, if we opt to say that the judgment or decision isn't either one of *those* events, but rather the more extended event that also includes the subsequent System 1 practical reasoning, then *that* event isn't an introspectable one. So either way, there is no one event, here, that is both introspectable and is a judgment/decision.

However, let me emphasize that the introspectable events that are involved in System 2 processes are by no means epiphenomenal. On the contrary. Nor, I should stress, is metacognition itself epiphenomenal either, contrary to a claim **Buckner et al.** make about the commitments of the target article. Quite the reverse. System 2 reasoning processes are shot through with – and are largely dependent upon – metacognitive thoughts and beliefs. And on any account, System 2 plays an important part in human cognition and behavior (albeit one that is subject to significant individual differences; see Stanovich 1999).

R10. The evidence from autism

The target article maintains that there is no convincing evidence that in autistic subjects metacognition is preserved while mindreading is damaged (sect. 10). This is contrary to the claims of Goldman (2006) and Nichols and Stich (2003), who cite such evidence in support of a "metacognition is prior" account, or a "two independent systems" view, respectively. **Williams, Lind, & Happé** (**Williams et al.**) agree with the target article in this respect, and cite Williams and Happé (in press a) as demonstrating that autistic children have equivalent difficulty attributing intentions to themselves and to other people, with their performance on these tasks being significantly correlated with their performance on traditional false-belief tasks. These new results are very welcome.

However, **Williams et al.** also cite evidence provided by Williams and Happé (in press b), which is said to favor a "one system, two modes of access" account over my preferred "mindreading is prior" thesis. In a modified version of the Smarties task, autistic children have significantly *greater* difficulty with the "self" version of the task than they do with the "other" version.[4] Williams et al. are mistaken in their interpretation of the significance of their own data, however. This is surprising, since all of the materials for a correct analysis are contained in the very article that they cite (Williams & Happé, in press b), as I shall now explain. The upshot is that these new data are fully consistent with a "mindreading is prior" account.

Suppose, first, that autistic children lack the normal mentalizing system altogether. (This seems to be the preferred view of Williams & Happé, in press b.) Such children would therefore lack whatever basic "core knowledge," or innate module, or innate domain-specific learning mechanism underlies the development of mentalizing abilities in normal children. Autistic children may nevertheless achieve some limited success in performance by other routes – by means of explicit domain-general theorizing, by memorizing rules and explicit strategies, and so forth. If an account of this sort is correct, then data from autistic subjects are inherently incapable of discriminating between the "mindreading is prior" and the "one mechanism, two modes of access" views of the relationship between mindreading and metacognition. For each of the latter applies only to those people who possess a normal (or near-normal) mentalizing *system*, or *faculty*. The "mindreading is prior" account claims that there is just a single mentalizing system, designed initially for mindreading, which is turned upon the self to issue in metacognition. In contrast, the "one mechanism, two modes of access" account, although agreeing that there is just a single mentalizing system, claims that the system in question has both perception-based and introspective channels of access to the mental items in its domain. The former predicts that no one with a normal mentalizing system should possess mindreading competence but lack metacognitive competence; whereas the latter predicts that there might be individuals with a normal mentalizing system who can mindread successfully but who lack a capacity for metacognition, because the introspective channel has been broken or disrupted. Importantly, *neither* model makes any predictions about what might happen in individuals who lack the normal mentalizing system altogether, but who rather "hack" their way to success by other methods. There might be all kinds of reasons why it could be easier to develop rules and strategies that apply to other people than it is to acquire such rules to apply to oneself, as Williams and Happé (in press b) themselves argue.

Now suppose, in contrast, that autistic children (or at least those who are comparatively high functioning) do possess a mentalizing system, only one that is significantly delayed in its normal development (and is perhaps slower and less reliable in its operations thereafter). And suppose that the "mindreading is prior" account of that system is correct. Still there might be reasons why individuals with a partly formed mentalizing faculty should find some mindreading tasks easier than parallel metacognitive ones. For example, as Williams and Happé (in press b)

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

themselves suggest, it may be that the perceptions of action that provide the main input for mentalizing are much more salient and easily accessible in the case of others' actions than in the case of one's own actions.[5]

However, wouldn't such an account predict (contrary to fact) that normally developing children should likewise pass mindreading tasks before they pass the equivalent metacognitive ones? Not necessarily. For the recent data mentioned in R3 suggest that a basic mentalizing *competence* is in place well before children start to be able to pass verbal versions of mentalizing tasks, and that there is some extraneous factor or factors that inhibit verbal performance. And the latter might contain no bias in favor of "other" versus "self" versions of the tasks. In the case of autistic children, in contrast, it is the delayed development of the mentalizing system itself that delays successful performance, enabling a bias in favor of mindreading over metacognition to display itself.

R11. Neuroimaging evidence

Lombardo, Chakrabarti, & Baron-Cohen (**Lombardo et al.**) cite neuroimaging evidence showing that identical neural regions are implicated in mentalizing about self and other, and that there are no other areas of the brain that are recruited specifically for mentalizing about self, or about other. These data are very welcome, and provide strong support for the "mindreading is prior" model. This is because all three of the competing accounts predict that there should be some brain regions used specifically for mentalizing about oneself and/or brain regions used specifically for mentalizing about others. Lombardo et al. claim, however, that it is an implication of the "mindreading is prior" account that the various brain regions implicated in mentalizing should be activated *to the same degree* when mentalizing about the self or about another. Because their data conflict with this prediction, they take this to raise a puzzle for the view. However, the "mindreading is prior" account makes no such prediction. For it allows that different kinds of data are implicated in the two forms of mentalizing. Specifically, mentalizing about the self can utilize visual and auditory imagery, somatosensory experiences, and so forth, in a way that mentalizing about others normally cannot. I suggest that these differences are sufficient to explain the different degrees of neural activation in question.

Nor, it should be stressed, does the "mindreading is prior" account predict that mindreading tasks are always performed in the same way (on the contrary; see sect. R2). So the findings reported by **Lombardo et al.** – that people tend to rely more on stereotypes when reasoning about the mental states of dissimilar others, while using simulation strategies when reasoning about the mental states of people who are perceived to be similar to themselves – raise no particular challenge for the account.

R12. The evidence from schizophrenia

The target article discusses the significance of the finding that schizophrenic patients with "passivity" symptoms have difficulties in attributing intentions to themselves while being normal at reading the minds of others. Nichols and Stich (2003) argue that this reveals a dissociation between metacognitive and mindreading abilities, whereas the target article suggests that the data are better explained in terms of *faulty* or *unusual* experiences being presented as input to an intact mindreading system. In contrast, **Wiffen & David** cast doubt upon the reliability of the data in question. If they are right, then that makes a "mindreading is prior" account even easier to defend.

Robbins, on the other hand, argues that schizophrenic patients with paranoid symptoms seem to display the contrary dissociation. For such patients perform poorly in mindreading tasks of various sorts, whereas there is no evidence (he tells us) that they show equivalent metacognitive deficits. **Wiffen & David** present two such strands of evidence, however. One is that schizophrenic patients characteristically lack insight into their own condition, which is (Wiffen & David claim) a failure of metacognition. But here they weave a tangled story. For although most if not all schizophrenic patients *do* perform poorly on tests of mindreading and *do* lack insight into their own illness, they appear to have no difficulties in distinguishing between normal and psychotic thoughts, feelings, and behavior in another person (Startup 1997). This raises a puzzle. If the lack of insight that these patients have into their own condition results from poor mindreading abilities, then how is it that they can nevertheless possess insight into the disordered minds of others?

We can begin to unravel this puzzle by noting that even if paranoid beliefs result partly from faulty mindreading, they cannot result from faulty mindreading *alone*. There must also exist a willingness to believe propositions whose prior probability is very low, in some circumstances. (Most of us may have entertained a paranoid *thought* or *hypothesis* at one time or another, but have immediately dismissed the idea as absurd.) And indeed, there is an extensive body of literature demonstrating that people with schizophrenia display a marked "jumping to conclusions" bias, forming beliefs from new data much more swiftly and with higher degrees of confidence than do controls. (See Blackwood et al. 2001, for a review.) Moreover, the bias in question seems to be one of data-gathering rather than a failure of probabilistic reasoning as such, since patients with schizophrenia reason normally about the plausibility of hypotheses that are presented to them, or when presented with the same range of data that lead normal individuals to formulate a new belief. This could explain why patients with schizophrenia lack insight into their own condition while showing insight into the conditions of others. For in the first case they are *forming* a paranoid belief from limited data, whereas in the latter case they are assessing the prior probability of someone else's belief.

Wiffen & David's other strand of evidence suggesting that schizophrenic patients have parallel difficulties in mindreading and metacognition is much more direct. They cite the demonstration by Koren et al. (2004) that schizophrenic patients do poorly on tests of metacognitive ability (see also Koren et al. 2006). Specifically, Koren et al. administered the Wisconsin Card Sorting Test to patients with schizophrenia, while also asking them to provide confidence ratings of their recent choices and while allowing them to decide whether or not each

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

sorting of the cards should count toward their final score (and potential monetary gain). The patients performed normally on the test itself, but displayed a marked deficit on the metacognitive measures. Although this is not yet an extensive body of data, it does suggest that deficits in mindreading and metacognition are paired together in schizophrenia, just as the "mindreading is prior" account would predict.

R13. Some friendly suggestions

A number of commentaries are entirely friendly to the approach taken in the target article, and hence need only a brief mention. First, **Evans** uses the position defended in the target article to resolve a tension in many theorists' thinking about dual systems of reasoning. For System 2 is often characterized as a conscious system, whereas we know that people's reports of System 2 processes are often confabulated. The solution is to note that only the globally broadcast contents of working memory are ever accessible to the mindreading system that is responsible for self-report, whereas many other aspects of System 2 processing will remain inaccessible to it. The contents of working memory represent but small islands of consciousness within the overall operations of System 2, leaving plenty of scope for confabulation about the remainder.

Second, **Huebner & Dennett** emphasize the dangers inherent in the use that is made of first-person pronouns throughout the target article, as in, "I have access to my own visual images," or, "We do have introspective access to inner speech." For these seem to imply a place for the *self* in the account, in addition to the various subpersonal systems described (for language, for mindreading, and so forth). Of course I intend no such thing. The outputs of the mindreading system are passed along as input to a variety of other systems, included in which is a language production mechanism that might issue in a (covert or overt) expression of the metarepresentational content in question; that is all. While use of personal pronouns in cognitive science is a handy *façon de parler*, we need to take care that their use is eliminable from the theories in question. I have no doubt, however, that they can be eliminated from all aspects of the "mindreading is prior" account.

Third, **Langland-Hassan** offers a welcome corrective to what I actually wrote in the target article, though not to anything that I believe or really intended to say. I had claimed that perceptual and quasi-perceptual states can be self-ascribed without interpretation by virtue of being globally broadcast. But Langland-Hassan points out that the question whether the speech that I seem to hear running through my head is my own or is really the voice of another person, cannot be answered without interpretation. For by hypothesis the mindreading system has no access to my own articulatory intentions. All it has access to is the resulting experience. Likewise for the question whether a visual image that I am currently entertaining is a memory-image or a fantasy-image. No experience can wear its own provenance on its face. Hence, describing myself as *remembering* the event depicted will have to be based on an inference grounded in aspects of the immediate context, feelings of familiarity, and so forth. All of this is entirely correct. What I should have said is that the *contents* of globally broadcast states can be self-attributed without interpretation, but interpretation is required for one to know to what *kind* those states belong. This leaves untouched the claim that the mindreading system has accessible to it data which it can use when self-ascribing propositional attitude states that are of no help in ascribing such states to other people.

R14. Behaviorism bites back

Catania offers a behaviorist alternative to my account, citing the work of Skinner (1945; 1963). Likewise, **Lewis & Carpendale** challenge the computationalist assumptions made by the target article, while criticizing me for not taking account of the work of the later Wittgenstein. I don't believe that I should need to argue in support of either cognitivism or computationalism, since both are foundational assumptions of most of cognitive science. In any case I don't have the space to defend them here. (See Gallistel & King [2009] for the definitive argument.) In addition, I don't believe that Wittgenstein's work contains any challenges that cognitive science cannot easily answer. There is some irony, moreover, in the charge that I should have paid more attention to Wittgenstein. For I spent the first fifteen years of my academic career focused on his philosophy, and much of that time was devoted to the so-called private language argument that Lewis & Carpendale refer to admiringly. This formed the topic of my doctoral dissertation. I ultimately came to believe that no version of the argument can be successful that doesn't already rely on anti-realist (e.g., behaviorist) or verificationist premises.

R15. Conclusion

I am grateful to my commentators for the care and attention they devoted to the target article. As a result, the theoretical options have been further clarified, and the "mindreading is prior" model of self-awareness has been additionally elaborated and strengthened. At the very least, that model will now need to be taken seriously by anyone considering the nature of self-awareness and its relationship to our mindreading abilities. And now that the strengths and weaknesses of the four main theoretical options have been clearly laid out, there is an urgent need for additional experimental data that will enable us to discriminate between them. As things stand, my own verdict is that the "mindreading is prior" account is the one that is best supported by the existing evidence (in part because it is the most parsimonious). But future findings could change all that.

NOTES

1. And then to respond to **Lurz**'s question why we should not believe that thoughts as well as perceptual states can be globally broadcast – raised also by **Pereplyotchik** – note that all of the evidence we have of global broadcasting concerns perceptual or quasi-perceptual events. And note, too, that the best established models of general-purpose working memory require the operation of one or another perceptual "slave system" – either the phonological loop or the visuospatial sketch pad; see Baddeley (1990).

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

2. Although **Kornell, Schwarz, & Son** (**Kornell et al.**) make a similar claim, in their case it is based on a misreading of my own view. So far as I can tell, they mean by "metacognition" precisely what I do.

3. Note that the behavior-guidance account of representational content proposed by **Anderson & Perlis** will also have exactly this consequence, because epistemic feelings guide action targeted on the world rather than on the animal's own mental states.

4. Note that this result is actually the reverse of the claims made by Goldman (2006) and Nichols and Stich (2003). For the data seem to show mindreading relatively intact while metacognition is damaged. **Lombardo et al.** mention similar data in respect of emotion understanding, showing that autistic people do significantly worse on measures of understanding their own emotions than they do on measures of understanding the emotions of others.

5. Note, too, that explanations similar to those provided here can accommodate the data cited by **Lombardo et al.** on autistic people's differential understanding of emotion in self and other.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Aiello, L. & Wheeler, P. (1995) The expensive tissue hypothesis. *Current Anthropology* 36:199–221. [aPC]

Al-Namlah, A. S., Fernyhough, C. & Meins, E. (2006) Sociocultural influences on the development of verbal mediation: Private speech and phonological recoding in Saudi Arabian and British samples. *Developmental Psychology* 42:117–31. [CF]

Ames, D. R. (2004) Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology* 87:340–53. [MVL]

Amsterdam, B. (1972) Mirror self-image reactions before age two. *Developmental Psychobiology* 5:297–305. [PRoc]

Anderson, J. (1995) *Learning and memory: An integrated approach*. Wiley. [aPC]

Anderson, J. R., Montant, M. & Schmitt, D. (1996) Rhesus monkeys fail to use gaze direction as an experimenter-given cue in an object-choice task. *Behavioural Processes* 37:47–55. [NK]

Anderson, M. L. & Oates, T. (2007) A review of recent research in metareasoning and metalearning. *AI Magazine* 28(1):7–16. [MLA]

Anderson, M. L., Oates, T., Chong, W. & Perlis, D. (2006) The metacognitive loop, I: Enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance. *Journal of Experimental and Theoretical Artificial Intelligence* 18(3):387–411. [MLA]

Anderson, M. L. & Perlis, D. (2005a) Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness. *Journal of Logic and Computation* 15:21–40. [MLA, aPC]

(2005b) The roots of self-awareness. *Phenomenology and the Cognitive Sciences* 4(3):297–333. [MLA]

Anderson, M. L. & Rosenberg, G. (2008) Content and action: The guidance theory of representation. *Journal of Mind and Behavior* 29(1–2):55–86. [MLA]

Andrade, J. (2001) The contribution of working memory to conscious experience. In: *Working memory in perspective*, ed. J. Andrade, pp. 60–79. Psychology Press. [JStBTE]

Apperly, I., Riggs, K., Simpson, A., Chiavarino, C. & Samson, D. (2007) Is belief reasoning automatic? *Psychological Science* 17:841–44. [rPC]

Baars, B. J. (1988) *A cognitive theory of consciousness*. Cambridge University Press. [aPC]

(1997) *In the theatre of consciousness*. Oxford University Press. [aPC, RWL]

(2002) The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Science* 6:47–52. [aPC]

(2003) How brain reveals mind: Neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies* 10:100–14. [aPC]

(2005) Subjective experience is probably not limited to humans: The evidence from neurobiology and behavior. *Consciousness and Cognition* 14:7–21. [BJB]

Baars, B. J. & Gage, N. M., eds. (2007) *Cognition, brain and consciousness: An introduction to cognitive neuroscience.* Elsevier/Academic Press. [BJB]

Baars, B. J., Ramsoy, T. & Laureys, S. (2003) Brain, consciousness, and the observing self. *Trends in Neurosciences* 26:671–75. [aPC]

Bacon, E., Izaute, M. & Danion, J. M. (2007) Preserved memory monitoring but impaired memory control during episodic encoding in patients with schizophrenia. *Journal of the International Neuropsychological Society* 13:219–27. [JP]

Baddeley, A. (1990) *Human memory*. Erlbaum. [rPC]

Baddeley, A., Chincotta, D. & Adlam, A. (2001) Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General* 130:641–57. [aPC]

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

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. (2001) The "Reading the Mind in the Eyes" test revised version: A study with normal adults and adults with Asperger syndrome and high-functioning autism. *Journal of Child Psychology and Psychiatry* 42:241–51. [MVL]

Bayne, T. & Pacherie, E. (2007) Narrators and comparators: The architecture of agentive self-awareness. *Synthese* 159:475–91. [aPC]

Beck, A. T., Baruch, E., Balter, J. M., Steer, R. A. & Warman, D. M. (2004) A new instrument for measuring insight: The Beck Cognitive Insight Scale. *Schizophrenia Research* 68(2–3):319–29. [BW]

Bedford, N. & David, A. (2008) Denial of illness in schizophrenia: Genuine or motivated? Unpublished doctoral dissertation, Institute of Psychiatry, King's College London. [BW]

Begg, I., Duft, S., Lalonde, P., Melnick, R. & Sanvito, J. (1989) Memory predictions are based on ease of processing. *Journal of Memory and Language* 28:610–32. [aPC]

Bem, D. (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74:183–200. [aPC]

(1972) Self-perception theory. In: *Advances in experimental social psychology, vol. 6*, ed. L. Berkowitz. Academic Press. [aPC]

Benjamin, A. & Bjork, R. (1996) Retrieval fluency as a metacognitive index. In: *Implicit memory and metacognition*, ed. L. Reder. Erlbaum. [aPC]

Beran, M. J., Smith, J. D., Coutinho, M. V. C., Couchman, J. J. & Boomer, J. G. (in press) The psychological organization of "uncertainty" responses and "middle" responses: A dissociation in capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes*. [JJC, rPC]

Beran, M., Smith, J., Redford, J. & Washburn, D. (2006) Rhesus macaques (*Macaca mulatta*) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes* 32:111–19. [aPC]

Bickhard, M. H. (2001) Why children don't have to solve the frame problems: Cognitive representations are not encodings. *Developmental Review* 21:224–62. [CL]

Bigelow, A. & Rochat, P. (2006) Two-month-old infants' sensitivity to social contingency in mother-infant and stranger-infant interaction. *Infancy* 9(3): 313–25. [PRoc]

Birch, S. A. J. & Bloom, P. (2004) Understanding children's and adult's limitations in mental state reasoning. *Trends in Cognitive Science* 8:255–60. [aPC]

(2007) The curse of knowledge in reasoning about false beliefs. *Psychological Science* 18:382–86. [aPC, MH]

Birch, S. A. J., Vauthier, S. A. & Bloom, P. (2008) Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition* 107:1018–34. [CMM]

Blackwood, N. J., Howard, R. J., Bentall, R. P. & Murray, R. M. (2001) Cognitive neuropsychiatric models of persecutory delusions. *American Journal of Psychiatry* 158:527–39. [PRob]

Blakemore, S., Wolpert, D. & Frith, C. (1998) Central cancellation of self-produced tickle sensation. *Nature Neuroscience* 1:635–40. [aPC]

Block, N. (1986) An advertisement for a semantics for psychology. In: *Midwest studies in philosophy: Vol. X, Studies in the philosophy of mind*, ed. P. French, T. Euhling & H. Wettstein. University of Minnesota Press. [rPC]

(1995) A confusion about the function of consciousness. *Behavioral and Brain Sciences* 18:227–47. [aPC]

Bloom, P. (2002) *How children learn the meaning of words*. MIT Press. [rPC]

(2005) *Descartes' baby.* Basic Books. [BH]

Bloom, P. & German, T. P. (2000) Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77:B25–B31. [CMM]

Bosco, F., Friedman, O. & Leslie, A. (2006) Recognition of pretend and real actions in play by 1- and 2-year-olds: Early success and why they fail. *Cognitive Development* 21:3–10. [arPC]

Botterill, G. & Carruthers, P. (1999) *The philosophy of psychology*. Cambridge University Press. [rPC]

Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L. & Hallett, M. (1992) Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry* 55:964–66. [aPC]

Bratman, M. (1987) *Intentions, plans, and practical reason*. Harvard University Press. [aPC]

(1999) *Faces of intention: Selected essays on intention and agency*. Cambridge University Press. [aPC]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Briñol, P. & Petty, R. (2003) Overt head movements and persuasion: A self-validation analysis. *Journal of Personality and Social Psychology* 84:1123–39. [aPC]

Brooks, R. & Meltzoff, A. N. (2002) The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology* 38:958–66. [MH]

Brüne, M. (2005) "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin* 31:21–42. [PRob]

Brüne, M., Lissek, S., Fuchs, N., Witthaus, H., Peters, S., Nicolas, V., Juckel, G. & Tegenthoff, M. (2008) An fMRI study of theory of mind in schizophrenic patients with "passivity" symptoms. *Neuropsychologia* 46(7):1992–2001. [BW]

Bruner, J. (1986) *Actual minds, possible worlds*. Harvard University Press. [CL]

(1990) *Acts of meaning*. Harvard University Press. [CL]

Brunet-Gouet, E. & Decety, J. (2006) Social brain dysfunctions in schizophrenia: A review of neuroimaging studies. *Psychiatry Research: Neuroimaging* 148:75–92. [PRob]

Byrne, R. & Whiten, A., eds. (1988) *Machiavellian intelligence I: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press. [aPC]

(1997) *Machiavellian intelligence II: Extensions and evaluations*. Cambridge University Press. [aPC]

Call, J. & Carpenter, M. (2001) Do apes and children know what they have seen? *Animal Cognition* 4:207–20. [aPC]

Call, J. & Tomasello, M. (2008) Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences* 12:187–92. [aPC]

Campbell, R. L. & Bickhard, M. H. (1993) Knowing levels and the child's understanding of mind. *Behavioral and Brain Sciences* 16:33–34. [CL]

Carpendale, J. I. M. & Chandler, M. J. (1996) On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development* 67:1686–1706. [CMM]

Carpendale, J. I. M. & Lewis, C. (2004) Constructing an understanding of mind: The development of children's social understanding within social interaction. *Behavioral and Brain Sciences* 27:79–151. [CL]

Carruthers, P. (1996a) Autism as mind-blindness. In: *Theories of Theories of Mind*, ed. P. Carruthers & P. Smith. Cambridge University Press. [aPC]

(1996b) *Language, thought and consciousness*. Cambridge University Press. [RTH]

(1996c) Simulation and self-knowledge. In: *Theories of Theories of Mind*, ed. P. Carruthers & P. Smith. Cambridge University Press. [aPC]

(2000) *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press. [aPC]

(2002) The cognitive functions of language. *Behavioral and Brain Sciences* 25:657–719. [aPC]

(2005) Why the question of animal consciousness might not matter very much. *Philosophical Psychology* 18:83–102. [CB]

(2006) *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford University Press. [arPC, JStBTE, KF]

(2008a) Cartesian epistemology: Is the theory of the self-transparent mind innate? *Journal of Consciousness Studies* 15(4):28–53. [aPC]

(2008b) Metacognition in animals: A skeptical look. *Mind and Language* 23(1):58–89. [CB, arPC, JJC, JP]

(2009) An architecture for dual reasoning. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 109–27. Oxford University Press. [arPC, KF, BH]

Catania, A. C. (1993) What John B. Watson left out of his behaviorism. *Mexican Journal of Behavior Analysis* 19:133–46. [ACC]

(2006) *Learning* (interim 4th edition). Sloan. [ACC]

Cheney, D. & Seyfarth, R. (2007) *Baboon metaphysics: The evolution of a social mind*. University of Chicago Press. [aPC]

Choe, K. S., Keil, F. C. & Bloom, P. (2005) Children's understanding of the Ulysses conflict. *Developmental Science* 8:387–92. [CMM]

Choi-Kain, L. W. & Gunderson, J. G. (2008) Mentalization: Ontogeny, assessment, and application in the treatment of borderline personality disorder. *American Journal of Psychiatry* 165:1127–35. [MH]

Clark, A. (1998) Magic words: How language augments human computation. In: *Language and thought*, ed. P. Carruthers & J. Boucher. Cambridge University Press. [aPC]

Cooper, J. & Duncan, B. (1971) Cognitive dissonance as a function of self-esteem and logical inconsistency. *Journal of Personality* 18:354–63. [aPC]

Corcoran, R. (2000) Theory of mind in other clinical conditions: Is a selective "theory of mind" deficit exclusive to autism? In: *Understanding other minds: Perspectives from developmental cognitive neuroscience*, 2nd edition, ed. S. Baron-Cohen, H. Tager-Flusberg & D. Cohen. Oxford University Press. [PRob]

Corcoran, R., Mercer, G. & Frith, C. D. (1995) Schizophrenia, symptomatology and social inference: investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research* 17(1):5–13. [BW]

Couchman, J. J., Coutinho, M. V. C., Beran, M. J. & Smith, J. D. (submitted) Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. [JJC]

Csibra, G. (2007) Action mirroring and action interpretation: An alternative account. In: *Sensorimotor foundations of higher cognition: Attention and performance XXII*, ed. P. Haggard, Y. Rosetti & M. Kawato. Oxford University Press. [aPC]

Csibra, G., Bíró, S., Koós, O. & Gergely, G. (2003) One-year-old infants use teleological representations of actions productively. *Cognitive Science* 27:111–33. [rPC]

Csibra, G. & Gergely, G. (2006) Social learning and social cognition: The case for pedagogy. In: *Processes of change in brain and cognitive development. Attention and performance XXI*, ed. Y. Munakata & M. H. Johnson, pp. 249–74. Oxford University Press. [MH]

Damasio, A. (1994) *Descartes' error: Emotion, reason and the human brain*. Papermac. [aPC]

(2003) *Looking for Spinoza: Joy, sorrow, and the feeling brain*. Harcourt. [aPC]

David, A. S. (2004) The clinical importance of insight: An overview. In: *Insight and psychosis: Awareness of illness in schizophrenia and related disorders*, 2nd edition, ed. X. F. Amador & A. S. David. Oxford University Press. [BW]

Davidoff, S. A., Forester, B. P., Ghaemi, S. N. & Bodkin, J. A. (1998) Effect of video self-observation on development of insight in psychotic disorders. *Journal of Nervous and Mental Disease* 186(11):697–700. [BW]

Decety, J. & Lamm, C. (2007) The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist* 13:580–93. [PL-H]

Dehaene, S. & Naccache, L. (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79:1–37. [aPC]

Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J. & Riviere, D. (2001) Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience* 4:752–58. [aPC]

Dehaene, S., Sergent, C. & Changeux, J. (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science* 100:8520–25. [aPC]

Dennett, D. C. (1978) Mechanism and responsibility. In: *Brainstorms.* MIT Press. [BH]

(1983) Intentional systems in cognitive ethology: The Panglossian paradigm defended. *Behavioral and Brain Sciences* 6(3):343–90. [CB]

(1991) *Consciousness explained*. Penguin. [aPC]

(2000) Making tools for thinking. In: *Metarepresentations*, ed. D. Sperber. Oxford University Press. [aPC]

Descartes, R. (1637/1985) *Discourse on the method*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch, *The Philosophical Writings of Descartes*, 2 vols. Cambridge University Press. (Original published in 1637.) [rPC]

Dewey, J. (1934/1980) *Art as experience*. Perigee Books. (Original work published in 1934.) [JJC]

Diamond, D., Stovall-McClough, C., Clarkin, J. F. & Levy, K. N. (2003) Patient-therapist attachment in the treatment of borderline personality disorder. *Bulletin of the Menninger Clinic* 67(3):227–59. [MH]

Dunbar, R. (2000) On the origin of the human mind. In: *Evolution and the human mind*, ed. P. Carruthers & A. Chamberlain. Cambridge University Press. [aPC]

Dunlosky, J. (2004) Metacognition. In: *Fundamentals of cognitive psychology*, 7th edition, ed. R. R. Hunt & H. C. Ellis. McGraw-Hill College. [MLA]

Dunlosky, J. & Bjork, R.A. (2008) *Handbook of memory and metamemory*. Psychology Press. [MLA]

Dunlosky, J. & Metcalfe, J. (2009) *Metacognition*. Sage. [MLA, rPC]

Eagly, A. & Chaiken, S. (1993) *The psychology of attitudes*. Harcourt Brace Jovanovich. [aPC]

Edwards, G. (1965) Post-hypnotic amnesia and post-hypnotic effect. *British Journal of Psychiatry* 111:316–25. [aPC]

Eichenbaum, N. J. & Cohen, N. J. (2001) *From conditioning to conscious reflection: Memory systems of the brain*. Oxford University Press. [JStBTE]

Eisenberger, N. I. & Lieberman, M. D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Sciences* 8(7):294–300. [BJB]

Esbensen, B. M., Taylor, M. & Stoess, C. (1997) Children's behavioral understanding of knowledge acquisition. *Cognitive Development* 12:53–84. [CMM]

Evans, G. (1982) *The varieties of reference*. Oxford University Press/Clarendon Press. [aPC, DM]

Evans, J. St. B. T. (1980) Thinking: Experiential and information processing approaches. In: *Cognitive psychology: New directions*, ed. G. Claxton, pp. 275–99. Routledge. [JStBTE]

(1989) *Bias in human reasoning: Causes and consequences*. Erlbaum. [JStBTE]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

(2008) Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology* 59:255–78. [JStBTE]
(2009) How many dual-process theories do we need: One, two or many? In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, p. 33–54. Oxford University Press. [JStBTE]
Evans, J. St. B. T. & Over, D. (1996) *Rationality and reasoning*. Psychology Press. [aPC, JStBTE]
Farrant, A., Boucher, J. & Blades, M. (1999) Metamemory in children with autism. *Child Development* 70:107–31. [aPC, JP]
Fazio, R. H. (1995) Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In: *Attitude strength: Antecedents and consequences, vol. 4*, ed. R. E. Petty & J. A. Krosnick, pp. 247–82. Erlbaum. [REP]
Fernyhough, C., Bland, K. A., Meins, E. & Coltheart, M. (2007) Imaginary companions and young children's responses to ambiguous auditory stimuli: Implications for typical and atypical development. *Journal of Child Psychology and Psychiatry* 48:1094–101. [CF]
Festinger, L. (1957) *A theory of cognitive dissonance*. Stanford University Press. [aPC]
Fisher, N., Happé, F. & Dunn, J. (2005) The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *Journal of Child Psychology and Psychiatry* 46:409–19. [DMW]
Flavell, J. H. (1979) Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry. *American Psychologist* 34:906–11. [rPC]
Flavell, J. H., Friedrichs, A. G. & Hoyt, J. D. (1970) Developmental changes in memorization processes. *Cognitive Psychology* 1:324–40. [CMM]
Flavell, J. H., Green, F. & Flavell, E. (1993) Children's understanding of the stream of consciousness. *Child Development* 64:387–98. [CF]
(2000) Development of children's awareness of their own thoughts. *Journal of Cognition and Development* 1:97–112. [CF]
Fodor, J. (1990) *A theory of content and other essays*. MIT Press. [rPC]
(1992) A theory of the child's theory of mind. *Cognition* 44:283–96. [arPC, OF]
Fonagy, P. & Bateman, A. (2008) The development of borderline personality disorder – a mentalizing model. *Journal of Personality Disorders* 22(1):4–21. [MH]
Fonagy, P., Gergely, G., Jurist, E. & Target, M. (2002) *Affect regulation, mentalization and the development of the self*. Other Press. [MH]
Fonagy, P., Gergely, G. & Target, M. (2007) The parent-infant dyad and the construction of the subjective self. *Journal of Child Psychology and Psychiatry* 48(3–4):288–328. [MH]
Fonagy, P., Leigh, T., Steele, M., Steele, H., Kennedy, R., Mattoon, G., Target M. & Gerber, A. (1996) The relation of attachment status, psychiatric classification, and response to psychotherapy. *Journal of Consulting and Clinical Psychology* 64:22–31. [MH]
Frankish, K. (1998) Natural language and virtual belief. In: *Language and thought: Interdisciplinary themes*, ed. P. Carruthers & J. Boucher, pp. 248–69. Cambridge University Press. [KF]
(2004) *Mind and supermind*. Cambridge University Press. [aPC, KF]
(2009) Systems and levels: Dual-system theories and the personal-subpersonal distinction. In: *In two minds: Dual processes and beyond*, ed. J. St. B. T. Evans & K. Frankish, pp. 89–107. Oxford University Press. [KF]
Frith, C. D. (1992) *The cognitive neuropsychology of schizophrenia*. Psychology Press. [BW]
(2007) *Making up the mind; how the brain creates our mental world*. Blackwell. [AZ]
Frith, C. D., Blakemore, S. & Wolpert, D. (2000a) Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B* 355:1771–88. [aPC, PL-H]
(2000b) Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* 31:357–63. [aPC, PL-H]
Frith, U. & Happé, F. (1999) Theory of mind and self-consciousness: What is it like to be autistic? *Mind and Language* 14:1–22. [aPC]
Gallagher, S. (2001) The practice of mind: Theory, simulation, or primary interaction? *Journal of Consciousness Studies* 8(5–7):83–107. [CB, rPC]
(2004) Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, and Psychology* 11(3):199–217. [MLA, CB, rPC]
(2005) *How the body shapes the mind*. Oxford University Press. [MLA]
Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G. (1996) Action recognition in the premotor cortex. *Brain* 119:593–609. [aPC]
Gallese, V. & Goldman, A. (1998) Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 12:493–501. [aPC]
Gallistel, C. R. & King, A. (2009) *Memory and the computational brain*. Wiley-Blackwell. [rPC]
Gazzaniga, M. (1995) Consciousness and the cerebral hemispheres. In: *The cognitive neurosciences*, ed. M. Gazzaniga. MIT Press. [arPC, PL-H]
(1998) *The mind's past*. California University Press. [aPC]
(2000) Cerebral specialization and inter-hemispheric communication: Does the corpus callosum enable the human condition? *Brain* 123:1293–326. [aPC]
Gergely, G. & Unoka, Z. (2008) Attachment, affect-regulation and mentalization: The developmental origins of the representational affective self. In: *Social cognition and developmental psychopathology*, ed. C. Sharp, P. Fonagy & I. Goodyer. Oxford University Press. [MH]
Gergely, G. & Watson, J. S. (1996) The social biofeedback theory of parental affect-mirroring: The development of emotional self-awareness and self-control in infancy. *The International Journal of Psycho-Analysis* 77:1–31. [MH]
(1999) Early social-emotional development: Contingency perception and the social biofeedback model. In: *Early social cognition*, ed. P. Rochat, pp. 101–37. Erlbaum. [MH]
Gigerenzer, G., Todd, P. & the ABC Research Group. (1999) *Simple heuristics that make us smart*. Oxford University Press. [aPC]
Gilleen, J., Greenwood, K. & David, A. S. (in press) Anosognosia in schizophrenia and other neuropsychiatric disorders: similarities and differences. In: *Advances in the study of anosognosia*, ed. G. P. Prigatano. Oxford University Press. [BW]
Gluck, M., Mercado, E. & Myers, C. (2008) *Learning and memory: From brain to behavior*. Worth Press. [CB]
Gluck, M. & Myers, C. (2001) *Gateway to memory: An introduction to neural network models of the hippocampus and learning*. MIT Press. [CB]
Goldberg, B. (1991) Mechanism and meaning. In: *Investigating psychology: Sciences of the mind after Wittgenstein*, ed. J. Hyman, pp. 48–66. Routledge. [CL]
Goldman, A. (1993) The psychology of folk psychology. *Behavioral and Brain Sciences* 16:15–28. [aPC]
(2006) *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press. [arPC, PL-H, RWL]
Gomez, J. (1998) Some thoughts about the evolution of LADS, with special reference to TOM and SAM. In: *Language and thought*, ed. P. Carruthers & J. Boucher. Cambridge University Press. [aPC]
Gopnik, A. (1993) The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16:1–14. [aPC, CF]
Gopnik, A. & Meltzoff, A. (1994) Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and theory of mind research. In: *Self-awareness in animals and humans*, ed. S. Parker, R. Mitchell & M. Boccia. Cambridge University Press. [aPC]
(1997) *Words, thoughts, and theories*. MIT Press. [arPC]
Gordon, R. (1986) Folk psychology as simulation. *Mind and Language* 1:158–70. [aPC]
(1996) "Radical" simulationism. In: *Theories of Theories of Mind*, ed. P. Carruthers & P. Smith. Cambridge University Press. [aPC]
Greenbaum, C. & Zemach, M. (1972) Role-playing and change of attitude toward the police after a campus riot: Effects of situational demand and justification. *Human Relations* 25:87–99. [aPC]
Grush, R. (2004) The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27:377–442. [aPC]
Hampton, R. R. (2001) Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences USA* 98:5359–62. [aPC, NK]
(2005) Can Rhesus monkeys discriminate between remembering and forgetting? In: *The Missing Link in Cognition: Origins of Self-reflective Consciousness*, ed. H. Terrace & J. Metcalfe. Oxford University Press. [aPC]
Hampton, R. R., Zivin, A. & Murray, E. (2004) Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition* 7:239–46. [aPC]
Happé, F. (2003) Theory of mind and the self. *Annals of the New York Academy of Sciences* 1001:134–44. [aPC]
Hare, B. (2007) From nonhuman to human mind: What changed and why? *Current Directions in Psychological Science* 16:60–64. [aPC]
Hare, B., Call, J., Agnetta, B. & Tomasello, M. (2000) Chimpanzees know what conspecifics do and do not see. *Animal Behavior* 59:771–85. [aPC]
Hare, B., Call, J. & Tomasello, M. (2001) Do chimpanzees know what conspecifics know? *Animal Behavior* 61:139–51. [aPC]
Harrington, L., Siegert, R. & McClure, J. (2005) Theory of mind in schizophrenia: A critical review. *Cognitive Neuropsychiatry* 10(4):249–86. [BW]
Harris, P. L. (2002a) Checking our sources: The origins of trust in testimony. *Studies in History and Philosophy of Science* 33:315–33. [aPC]
(2002b) What do children learn from testimony? In: *The cognitive basis of science*, ed. P. Carruthers, S. Stich & M. Siegal. Cambridge University Press. [aPC]
(2007) Trust. *Developmental Science* 10:135–38. [CMM]
Haslam, N. (2006) Dehumanization: An integrative review. *Personality and Social Psychology* 10(3):252–64. [BH]
Heavey, C. L. & Hurlburt, R. T. (2008) The phenomena of inner experience. *Consciousness and Cognition* 17:798–810. [RTH]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Heil, J. (1981) Does cognitive psychology rest on a mistake? *Mind* 90:321–42. [CL]
Heyes, C. M. (1998) Theory of mind in nonhuman primates. *Behavioral and Brain Sciences* 21(1):101–34. [JJC]
Hobson, P. (2002) *The cradle of thought: Explorations of the origins of thinking*. Macmillan. [CL]
Hogrefe, G. J, Wimmer, H. & Perner, J. (1986) Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development* 57:567–82. [DMW]
Horowitz, A. (in press) Attention to attention in domestic dog (*Canis familiaris*) dyadic play. *Animal Cognition*. [NK]
Hurlburt, R. (1990) *Sampling normal and schizophrenic inner experience.* Plenum Press. [arPC, RTH]
(1993) *Sampling inner experience in disturbed affect*. Plenum Press. [arPC, RTH]
(1997) Randomly sampling thinking in the natural environment. *Journal of Consulting and Clinical Psychology* 65(6):941–49. [RTH, DM]
(2006) *Comprehending behavioral statistics*, 4th edition. Wadsworth. [RTH]
Hurlburt, R. T. & Akhter, S. A. (2006) The Descriptive Experience Sampling method. *Phenomenology and the Cognitive Sciences* 5:271–301. [RTH]
(2008) Unsymbolized thinking. *Consciousness and Cognition* 17:1364–74. [arPC, RTH]
Hurlburt, R. T., Happé, F. & Frith, U. (1994) Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine* 24:385–95. [aPC]
Hurlburt, R. T. & Heavey, C. L. (2006) *Exploring inner experience: The Descriptive Experience Sampling method*. John Benjamins. [CF, RTH, rPC]
(in preparation) Sensory awareness. [RTH]
Hurlburt, R. T. & Schwitzgebel, E. (2007) *Describing inner experience? Proponent meets skeptic*. Bradford Books/MIT Press. [CF, RTH]
Hutto, D. (2004) The limits of spectatorial folk psychology. *Mind and Language* 19:548–73. [CB]
(2008) *Folk psychological narratives.* MIT Press. [CB]
Hutton, R. & Sameth, J. (1988) *The mind. Part 1: The search for the mind* [Video]. Video film edited by R. Hutton; directed by J. Sameth. Annenberg/CPB Project. [BF]
Inman, A. & Shettleworth, S. J. (1999) Detecting metamemory in non-verbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* 25:389–95. [JJC, rPC]
Jackendoff, R. (1996) How language helps us think. *Pragmatics and Cognition* 4(1):1–34. [BH]
James, W. (1872) Are we automata? *Mind* 4:1–22. [PRoc]
(1890/1952) *The principles of psychology*. In series: *Great Books of the Western World, vol. 53*, ed. R. M. Hutchins. University of Chicago Press. (Original work published in 1890). [JJC]
Jaswal, V. K. & Neely, L. A. (2006) Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science* 17:757–58. [CMM]
Jeannerod, M. (2006) *Motor cognition*. Oxford University Press. [aPC]
Jeannerod, M. & Pacherie, E. (2004) Agency, simulation and self-identification. *Mind and Language* 19:113–46. [PL-H]
Jenkins, A. C., Macrae, C. N. & Mitchell, J. P. (2008) Repetition suppression of ventromedial prefrontal activity during judgments of self and other. *Proceedings of the National Academy of Sciences USA* 105:4507–12. [MVL]
Johnson, J. G., Cohen, P., Chen, H., Kasen, S. & Brook, J. S. (2006) Parenting behaviors associated with risk for offspring personality disorder during adulthood. *Archives of General Psychiatry* 63(5):579–87. [MH]
Johnson, S. (2000) The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences* 4:22–28. [rPC]
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. [aPC]
Kant, I. (1781/1929) *The critique of pure reason*, trans. N. Kemp Smith, Macmillan. (Original published in 1781.) [aPC]
Kazak, S., Collis, G. & Lewis, V. (1997) Can young people with autism refer to knowledge states? Evidence from their understanding of "know" and "guess." *Journal of Child Psychology and Psychiatry* 38:1001–1009. [aPC, PRob, DMW]
Kenny, A. (1991) The homunculus fallacy. In: *Investigating psychology: Sciences of the mind after Wittgenstein*, ed. J. Hyman, pp. 155–65. Routledge. (Original work published in 1971.) [CL]
Keysar, B., Lin, S. & Barr, D. J. (2003) Limits on theory of mind use in adults. *Cognition* 89:25–41. [rPC, MH]
Koenig, M. & Harris, P. L. (2005) Preschoolers mistrust ignorant and inaccurate speakers. *Child Development* 76(6):1261–77. [CMM]
Koren, D., Seidman, L. J., Goldsmith, M. & Harvey, P. D. (2006) Real-world cognitive – and metacognitive – dysfunction in schizophrenia: A new approach for measuring (and remediating) more "right stuff." *Schizophrenia Bulletin* 32(2):310–26. [JP, rPC]
Koren, D., Seidman, L. J., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S. & Klein E. (2004) The neuropsychological basis of insight in first-episode schizophrenia: A pilot metacognitive study. *Schizophrenia Research* 70(2–3):195–202. [BW, rPC]
Koriat, A. (1993) How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review* 100:609–39. [aPC]
(1997) Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General* 126:349–70. [aPC]
(2000) The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition* 9:149–71. [JP]
(2007) Metacognition and consciousness. In: *The Cambridge handbook of consciousness*, ed. P. Zelazo, M. Moscovitch & E. Thompson. Cambridge University Press. [rPC]
Koriat, A., Ma'ayan, H. & Nussinson, R. (2006) The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General* 135(1):36–69. [arPC, JP]
Kornell, N. (in press) Metacognition in humans and animals. *Current Directions in Psychological Science*. [NK]
Kornell, N., Son, L. & Terrace, H. (2007) Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science* 18:64–71. [arPC, NK]
Kosslyn, S. (1994) *Image and brain*. MIT Press. [aPC]
Kreiman, G., Fried, I. & Koch, C. (2003) Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Science* 99:8378–83. [aPC]
Kruglanski, A., Alon, W. & Lewis, T. (1972) Retrospective misattribution and task enjoyment. *Journal of Experimental Social Psychology* 8:493–501. [aPC]
Langland-Hassan, P. (2008) Fractured phenomenologies: Thought insertion, inner speech, and the puzzle of extraneity. *Mind and Language* 23:369–401. [PL-H]
Leslie, A. M. (1987) Pretense and representation: The origins of "theory of mind." *Psychological Review* 94:412–26. [MH]
Leslie, A. M., Friedman, O. & German, T. P. (2004) Core mechanisms in "theory of mind." *Trends in Cognitive Sciences* 8:528–33. [OF, rPC]
Leslie, A. M. & Polizzi, P. (1998) Inhibitory processing in the false belief task: Two conjectures. *Developmental Science* 1:247–53. [aPC]
Leslie, A. M. & Thaiss, L. (1992) Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition* 43:225–51. [OF]
Leudar, I. & Costall, A., eds. (in press) *Against theory of mind.* [CL]
Levelt, W. (1989) *Speaking: From intention to articulation*. MIT Press. [aPC]
Levy, K. N., Meehan, K. B., Kelly, K. M., Reynoso, J. S., Weber, M., Clarkin, J. F. & Kernberg O. F. (2006) Change in attachment patterns and reflective function in a randomized control trial of transference-focused psychotherapy for borderline personality disorder. *Journal of Consulting and Clinical Psychology* 74(6):1027–40. [MH]
Lewis, M. (1992) *Shame: The exposed self.* Free Press. [PRoc]
Lind, S. E. (2008) Episodic memory, "theory of mind," and temporally extended self-awareness in autism spectrum disorder. Unpublished doctoral dissertation, City University, London. [DMW]
Lind, S. E. & Bowler, D. M. (2008) Episodic memory and autonoetic consciousness in autism spectrum disorders: The roles of self-awareness, representational abilities and temporal cognition. In: *Memory in autism: Theory and evidence*, ed. J. M. Boucher & D. M. Bowler, pp. 166–87. Cambridge University Press. [DMW]
(under revised review) Self-other source memory and its relation to theory-of-mind in autism spectrum disorder. *Journal of Autism and Developmental Disorders*. [DMW]
Linehan, M. M. (1993) *Cognitive-behavioural treatment of borderline personality disorder*. Guilford Press. [MH]
Locke, J. (1690/1961) *An essay concerning human understanding*, ed. J. Yolton, Dent and Sons. (Original published in 1690.) [aPC]
Lombardo, M. V., Barnes, J. L., Wheelwright, S. J. & Baron-Cohen, S. (2007) Self-referential cognition and empathy in autism. *PLoS One* 2:e883. [MVL]
Lombardo, M. V., Chakrabarti, B. C., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J. S., Baron-Cohen, S. & MRC AIMS Consortium. (submitted) My connection with your mind: Identical functional connectivity from shared neural circuits for mentalizing about the self and others. [MVL]
Luo, Y. & Baillargeon, R. (2005) Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science* 16:601–608. [rPC]
Lutz, D. & Keil, F. (2002) Early understanding of the division of cognitive labor. *Child Development* 73:1073–84. [CMM]
Lyons-Ruth, K., Yellin, C., Melnick, S. & Atwood, G. (2005) Expanding the concept of unresolved mental states: Hostile/helpless states of mind on the Adult Attachment Interview are associated with disrupted mother-infant

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

communication and infant disorganization. *Developmental Psychopathology* 17(1):1–23. [MH]
Marcel, A. J., Tegnér, R. & Nimmo-Smith, I. (2004) Anosognosia for plegia: Specificity, extension, partiality and disunity of bodily unawareness. *Cortex* 40(1):19–40. [BW]
McCloskey, M. (1983) Naive theories of motion. In: *Mental models*, ed. D. Gentner & A. Stevens. Erlbaum. [aPC]
McEvoy, J. P., Schooler, N. R., Friedman, E., Steingard, S. & Allen, M. (1993) Use of psychopathology vignettes by patients with schizophrenia or schizoaffective disorder and by mental health professionals to judge patients' insight. *American Journal of Psychiatry* 150(11):1649–53. [BW]
McGeer, V. (2004) Autistic self-awareness. *Philosophy, Psychiatry, and Psychology* 11(3):235–51. [DMW]
Medalia, A. & Thysen, J. (2008) Insight into neurocognitive dysfunction in schizophrenia. *Schizophrenia Bulletin* 34(6):1221–30. [BW]
Meins, E., Fernyhough, C., Wainwright, R., Clark-Carter, D., Das Gupta, M., Fradley, E. & Tuckey, M. (2003) Pathways to understanding mind: Construct validity and predictive validity of maternal mind-mindedness. *Child Development* 74:1194–211. [CF]
Meltzoff, A. (1995) Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31:838–50. [CMM]
Metcalfe, J. (1993) Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review* 100:3–22. [MLA]
(2008) Evolution of metacognition. In: *Handbook of metacognition and learning*, ed. J. Dunlosky & R. Bjork. Erlbaum. [rPC]
Metcalfe, J. & Greene, M. J. (2007) Metacognition of agency. *Journal of Experimental Psychology General* 136(2):184–99. [JP]
Metcalfe, J. & Shimamura, A., eds. (1994) *Metacognition: Knowing about knowing*. MIT Press. [MLA, aPC]
Metzinger, T. (2004) *Being no one*. MIT Press. [BH]
Mills, C. M. & Keil, F. C. (2004) Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology* 87:1–32. [CMM]
Mitchell, J. P., Ames, D. L., Jenkins, A. C. & Banaji, M. R. (in press) Neural correlates of stereotype application. *Journal of Cognitive Neuroscience.* [MVL]
Mitchell, J. P., Macrae, C. N. & Banaji, M. R. (2006) Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* 50:655–63. [MVL]
Moll, H. & Tomasello, M. (2007) Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London* 362:639–48. [CL]
Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–50. [JJC]
Nelson, E. E. & Panksepp, J. (1998) Brain substrates of infant–mother attachment: Contributions of opioids, oxytocin, and norepinephrine. *Neuroscience and Biobehavioral Reviews* 22(3):437–52. [BJB]
Nelson, T. O, ed. (1992) *Metacognition: Core readings*. Allyn and Bacon. [aPC]
Nelson, T. O. & Narens, L. (1990) Metamemory: A theoretical framework and some new findings. In: *The psychology of learning and motivation, vol. 26*, ed. G. H. Bower, pp. 125–73. Academic Press. [MLA, NK]
Nichols, S. & Stich, S. (2003) *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press. [arPC, PL-H, PRob, DMW]
Nisbett, R. & Wilson, T. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–95. [aPC, BF, OF, REP]
Onishi, K. & Baillargeon, R. (2005) Do 15-month-olds understand false beliefs? *Science* 5719:255–58. [CB, arPC]
Onishi, K., Baillargeon, R. & Leslie, A. (2007) 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica* 124:106–28. [arPC, MH]
Origgi, G. & Sperber, D. (2000) Evolution, communication, and the proper function of language. In: *The evolution of the human mind*, ed. P. Carruthers & A. Chamberlain. Cambridge University Press. [aPC]
Paulescu, E., Frith, D. & Frackowiak, R. (1993) The neural correlates of the verbal component of working memory. *Nature* 362:342–45. [aPC]
Perlis, D. (1985) Languages with self-reference: I. Foundations. *Artificial Intelligence* 25:301–22. [MLA]
(1988) Languages with self-reference: II. Knowledge, belief, and modality. *Artificial Intelligence* 34:179–212. [MLA]
(1997) Consciousness as self-function. *Journal of Consciousness Studies* 4(5–6):509–25. [MLA]
(2000) What does it take to refer? *Journal of Consciousness Studies* 7(5):67–69. [MLA]
Perlis, D. & Subrahmanian, V. S. (1994) Metalanguages, reflection principles and self-reference. In: *Handbook of logic in artificial intelligence and logic programming, vol. 2: Deduction methodologies*, ed. D. Gabbay, C. J. Hogger & J. A. Robinson. Oxford University Press. [MLA]
Perner, J. (2000) Memory and theory of mind. In: *The Oxford handbook of memory*, ed. E. Tulving & F. I. M. Craik, pp. 297–312. Oxford University Press. [DMW]
Perner, J., Rendl, B. & Garnham, A. (2007) Objects of desire, thought, and reality: Problems of anchoring discourse referents in development. *Mind and Language* 22(5):475–513. [CB]
Perner, J. & Ruffman, T. (2005) Infants' insight into the mind: How deep? *Science* 308:214–16. [CB]
Pessiglione, M., Schmidt, L., Palminteri, S. & Frith, C. D. (in press) Reward processing and conscious awareness. In: *Attention and performance XX111*, ed. M. Delgado, E. A. Phelps & T. W. Robbins. Oxford University Press. [AZ]
Petty, R. E. & Cacioppo, J. T. (1986) *Communication and persuasion: Central and peripheral routes to attitude change.* Springer-Verlag. [REP]
Phillips, W., Baron-Cohen, S. & Rutter, M. (1998) Understanding intention in normal development and in autism. *British Journal of Developmental Psychology* 16:337–48. [aPC]
Pickup, G. J. & Frith, C. D. (2001) Theory of mind impairments in schizophrenia: Symptomatology, severity and specificity. *Psychological Medicine* 31(2):207–20. [BW]
Pousa, E., Duñó, R., Blas Navarro, J., Ruiz, A. I., Obiols, J. E. & David, A. S. (2008) Exploratory study of the association between insight and Theory of Mind (ToM) in stable schizophrenia patients. *Cognitive Neuropsychiatry* 13(3):210–32. [BW]
Proudfoot, D. (1997) On Wittgenstein on cognitive science. *Philosophy* 72:189–217. [CL]
Proust, J. (2007) Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese* 2:271–95. [JP, rPC]
(in press a) Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society* 108(Pt 3):241–68. [JP]
(in press b) The representational basis of brute metacognition: A proposal. In: *Philosophy of animal minds: New essays on animal thought and consciousness*, ed. R. Lurz. Cambridge University Press. [JP]
Putnam, K. M. & Silk, K. R. (2005) Emotion dysregulation and the development of borderline personality disorder. *Developmental Psychopathology* 17(4):899–925. [MH]
Raffman, D. (1999) What autism may tell us about self-awareness: A commentary on Frith and Happé. *Mind and Language* 14(1):23–31. [DMW]
Reddy, V. (2003) On being an object of attention: Implications for self-other-consciousness. *Trends in Cognitive Science* 7(9):397–402. [PRoc]
(2008) *How infants know minds*. Harvard University Press. [CL]
Reder, L. M., ed. (1996) *Implicit memory and metacognition.* Erlbaum. [NK]
Rey, G. (2008) (Even higher-order) intentionality without consciousness. *Revue Internationale de Philosophie* 62:51–78. [aPC]
Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996) Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3:131–41. [aPC]
Robbins, P. (2006) The ins and outs of introspection. *Philosophy Compass* 1:617–30. [PRob]
Robbins, P. & Jack, A. (2006) The phenomenal stance. *Philosophical Studies* 127:59–85. [BH]
Rochat, P. (2009) *Others in mind – Social origins of self-consciousness*. Cambridge University Press. [PRoc]
Rockeach, M. (1964) *The three Christs of Ypsilanti*. Knopf. [BW]
Rosenthal, D. M. (2005) *Consciousness and mind.* Clarendon Press. [DP]
Ruffman, T. & Perner, J. (2005) Do infants really understand false belief? Response to Leslie. *Trends in Cognitive Sciences* 9(10):462–63. [CB]
Russell, J. & Hill, E. (2001) Action-monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry* 42:317–28. [aPC]
Sabbagh, M. A. & Baldwin, D. A. (2001) Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development* 72:1054–70. [CMM]
Santos, L. R., Nissen, A. G. & Ferrugia, J. A. (2006) Rhesus monkeys, *Macaca mulatta*, know what others can and cannot hear. *Animal Behaviour* 71:1175–81. [NK]
Schmideberg, M. (1947) The treatment of psychopathic and borderline patients. *American Journal of Psychotherapy* 1:45–71. [MH]
Scholl, B. (2007) Object persistence in philosophy and psychology. *Mind and Language* 22:563–91. [rPC]
Schwartz, B. L., Benjamin, A. S. & Bjork, R. A. (1997) The inferential and experiential basis of metamemory. *Current Directions in Psychological Science* 6:132–37. [NK]
Schwartz, B. L. & Smith, S. (1997) The retrieval of related information influences tip-of-the-tongue states. *Journal of Memory and Language* 36:68–86. [aPC]
Schwarz, N. & Bohner, G. (2000) The construction of attitudes. In: *Blackwell handbook of social psychology: Intrapersonal processes*, ed. A. Tesser & N. Schwarz. Blackwell. [REP]
Searle, J. (1992) *The rediscovery of the mind*. MIT Press. [aPC]
Sellars, W. (1956/1997) *Empiricism and the philosophy of mind*. Harvard University Press. (Original work published in 1956). [DP]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Senju, A., Csibra, G. & Johnson, M. H. (2008) Understanding the referential nature of looking: Infants' preference for object-directed gaze. *Cognition* 108:303–19. [MH]

Shallice, T. (1988) *From neuropsychology to mental structure*. Cambridge University Press. [aPC]

Shanahan, M. & Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition* 98:157–76. [rPC]

Sharp, C. & Fonagy, P. (2008) The parent's capacity to treat the child as a psychological agent: Constructs, measures and implications for developmental psychopathology. *Social Development* 17(3):737–54. [MH]

Sheehan, P. & Orne, M. (1968) Some comments on the nature of post-hypnotic behavior. *Journal of Nervous and Mental Disease* 146:209–20. [aPC]

Shergill, S., Brammer, M., Fukuda, R., Bullmore, E., Amaro, E., Murray, R. & McGuire, P. (2002) Modulation of activity in temporal cortex during generation of inner speech. *Human Brain Mapping* 16:219–27. [aPC]

Shields, W., Smith, J. & Washburn, D. (1997) Uncertain responses by humans and rhesus monkeys (*Macaca mulatta*) in a psychophysical same–different task. *Journal of Experimental Psychology: General* 126:147–64. [aPC]

Shoemaker, S. (1996) *The first-person perspective and other essays*. Cambridge University Press. [aPC]

Shriver, A. & Allen, C. (2005) Consciousness might matter very much. *Philosophical Psychology* 18:103–11. [CB]

Siegal, M. & Varley, R. (2002) Neural systems involved in theory of mind. *Nature Reviews Neuroscience* 3:462–71. [rPC]

Siever, L. J., Torgersen, S., Gunderson, J. G., Livesley, W. J. & Kendler, K. S. (2002) The borderline diagnosis, III: Identifying endophenotypes for genetic studies. *Biological Psychiatry* 51(12):964–68. [MH]

Skinner, B. F. (1945) The operational analysis of psychological terms. *Psychological Review* 52:270–77. [ACC, rPC]

(1963) Behaviorism at fifty. *Science* 140:951–58. [ACC, rPC]

(1969) An operant analysis of problem solving. In: B. F. Skinner, *Contingencies of reinforcement*. Appleton-Century-Crofts. [ACC]

Sloman, S. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119:3–22. [aPC]

(2002) Two systems of reasoning. In: *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin & D. Kahneman. Cambridge University Press. [aPC]

Smith, E. R. & DeCoster, J. (2000) Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review* 4:108–31. [JStBTE]

Smith, J. D. (2005) Studies of uncertainty monitoring and metacognition in animals and humans. In: *The missing link in cognition: Origins of self-reflective consciousness*, ed. H. Terrace & J. Metcalfe. Oxford University Press. [aPC]

Smith, J. D., Beran, M. J., Couchman, J. J. & Coutinho, M. V. C. (2008) The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin and Review* 15(4):679–91. [JJC]

Smith, J. D., Beran, M. J., Redford, J. S. & Washburn, D. A. (2006) Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General* 135(2):282–97. [JJC, rPC]

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

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

Smith, J. D., Shields, W. & Washburn, D. (2003) The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences* 26:317–39; discussion pp. 339–73. [aPC, JJC, NK]

Smith, J. D. & Washburn, D. A. (2005) Uncertainty monitoring and metacognition by animals. *Current Directions in Psychological Science* 14:19–24. [NK]

Son, L. & Kornell, N. (2005) Meta-confidence judgments in rhesus macaques: Explicit versus implicit mechanisms. In: *The missing link in cognition: Origins of self-reflective consciousness*, ed. H. Terrace & J. Metcalfe. Oxford University Press. [arPC, NK]

Song, H. & Baillargeon, R. (forthcoming) Infants' reasoning about others' false perceptions. *Developmental Psychology*. [arPC]

Song, H., Onishi, K., Baillargeon, R. & Fisher, C. (forthcoming) Can an agent's false belief be corrected through an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*. [arPC]

Southgate, V., Gergely, G. & Csibra, G. (2008) Does the mirror neuron system and its impairment explain human imitation and autism? In: *The role of mirroring processes in social cognition*, ed. J. Pineda. Humana Press. [aPC]

Southgate, V., Senju, A. & Csibra, G. (2007) Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* 18:587–92. [arPC]

Sperber, D. & Wilson, D. (1995) *Relevance: Communication and cognition*, 2nd edition. Blackwell. [aPC]

Sprong, M., Schothorst, P., Vos, E., Hox, J. & Van Engeland, H. (2007) Theory of mind in schizophrenia: Meta-analysis. *British Journal of Psychiatry* 191(1):5–13. [BW]

Stanovich, K. (1999) *Who is rational? Studies of individual differences in reasoning*. Erlbaum. [arPC]

Startup, M. (1997) Awareness of own and others' schizophrenic illness. *Schizophrenia Research* 26(2–3):203–11. [BW, rPC]

Stenning, K. & Van Lambalgen, M. (2008) *Human reasoning and cognitive science*. MIT Press. [CB]

Sternberg, S. (2001) Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica* 106:147–246. [aPC]

Suda-King, C. (2008) Do orangutans (*Pongo pygmaeus*) know when they do not remember? *Animal Cognition* 11:21–42. [NK]

Surian, L., Caldi, S. & Sperber, D. (2007) Attribution of beliefs by 13-month old infants. *Psychological Science* 18:580–86. [CB, arPC]

Taylor, M., Esbensen, B. & Bennett, R. (1994) Children's understanding of knowledge acquisition: The tendency for children to report they have always known what they have just learned. *Child Development* 65:1581–604. [CMM]

Terrace, H. & Metcalfe, J., eds. (2005) *The missing link in cognition: Origins of self-reflective consciousness*. Oxford University Press. [aPC]

Tolman, E. C. (1938) The determiners of behavior at a choice point. *Psychological Review* 45:1–41. [JJC]

Tomasello, M. (2003) *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press. [CL]

Tomasello, M., Call, J. & Hare, B. (2003a) Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences* 7:153–56. [aPC]

(2003b) Chimpanzees versus humans: It's not that simple. *Trends in Cognitive Sciences* 7:239–40. [aPC]

Tomasello, M., Hare, B. & Agnetta, B. (1999) A nonverbal false belief task: The performance of children and great apes. *Child Development* 70:381–95. [NK]

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A. & Csibra, G. (2008) Infant perseverative errors are induced by pragmatic misinterpretation. *Science* 321:1831–34. [PRoc]

Tronick, E. Z. (2005) Why is connection with others so critical? The formation of dyadic states of consciousness and the expansion of individuals' states of consciousness: Coherence governed selection and the co-creation of meaning out of messy meaning making. In: *Emotional development*, ed. J. Nadel & D. Muir, pp. 293–315. Oxford University Press. [PRoc]

Tronick, E. Z., Als, H., Adamson, L., Wise, S. & Brazelton, T. B. (1978) The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry* 17:1–13. [PRoc]

Turnbull, W. (2003) *Language in action: Psychological models of conversation*. Psychology Press. [CL]

VanderBorght, M. & Jaswal, V. K. (in press) Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development*. [CMM]

Varley, R. (1998) Aphasic language, aphasic thought. In: *Language and thought*, ed. P. Carruthers & J. Boucher. Cambridge University Press. [rPC]

von Helmholtz, H. (1866) *Handbuch der Physiologischen Optik*. Voss. [AZ]

Vygotsky, L. S. (1934/1987) Thinking and speech. In: *The Collected Works of L. S. Vygotsky, vol. 1*. Plenum Press. (Original work published in 1934). [CF]

Warman, D. M., Lysaker, P. H. & Martin, J. M. (2007) Cognitive insight and psychotic disorder: The impact of active delusions. *Schizophrenia Research* 90(1–3):325–33. [BW]

Washburn, D., Smith, J. & Shields, W. (2006) Rhesus monkeys (*Macaca mulatta*) immediately generalize the *uncertain* response. *Journal of Experimental Psychology: Animal Behavior Processes* 32:185–89. [aPC]

Wason, P. C. & Evans, J. St. B. T. (1975) Dual processes in reasoning? *Cognition* 3:141–54. [JStBTE]

Wegner, D. (2002) *The illusion of conscious will*. MIT Press. [aPC]

Wegner, D. & Wheatley, T. (1999) Apparent mental causation: Sources of the experience of the will. *American Psychologist* 54:480–91. [aPC]

Weiskrantz, L., Elliot, J. & Darlington, C. (1971) Preliminary observations of tickling oneself. *Nature* 230:598–99. [aPC]

Wellman, H. (1990) *The child's theory of mind*. MIT Press. [arPC]

Wellman, H., Cross, D. & Watson, J. (2001) Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72:655–84. [aPC, CMM, PRob]

Wells, G. & Petty, R. (1980) The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology* 1:219–30. [aPC]

Wheatley, T. & Haidt, J. (2005) Hypnotic disgust makes moral judgments more severe. *Psychological Science* 16:780–84. [BF]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press

Wicklund, R. & Brehm, J. (1976) *Perspectives on cognitive dissonance*. Erlbaum. [aPC]

Williams, D. M. (2008) Conceptual and pre-conceptual awareness of self and other: Studies of autism and typical development. Unpublished doctoral dissertation, University of London. [DMW]

Williams, D. M. & Happé, F. (in press a) Representing intentions in self and other: Studies of autism and typical development. *Developmental Science*. [DMW, rPC]

(in press b) "What did I say?" versus "What did I think?": Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders*. [DMW, rPC]

Wilson, T. (2002) *Strangers to ourselves*. Harvard University Press. [aPC]

Wimmer, H., Hogrefe, G. & Perner, J. (1988) Children's understanding of informational access as a source of knowledge. *Child Development* 59:386–96. [PRob, rPC]

Winsler, A. & Naglieri, J. (2003) Overt and covert verbal problem-solving strategies: Developmental trends in use, awareness, and relations with task performance in children aged 5 to 17. *Child Development* 74:659–78. [CF]

Wittgenstein, L. (1968) *Philosophical investigations*. Blackwell. [CL]

Wixted, J. T. & Gaitan, S. C. (2002) Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning and Behavior* 30:289–305. [ACC]

Wolpert, D. & Ghahramani, Z. (2000) Computational principles of movement neuroscience. *Nature Neuroscience* 3:1212–17. [aPC]

Wolpert, D. & Kawato, M. (1998) Multiple paired forward and inverse models for motor control. *Neural Networks* 11:1317–29. [aPC]

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

Zelazo, P. D. (2004) The development of conscious control in childhood. *Trends in Cognitive Science* 8(1):12–17. [AZ]

https://doi.org/10.1017/S0140525X09000624 Published online by Cambridge University Press