Forthcoming as a target paper in *Behavioral and Brain Sciences*, 32 (2009)

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

Peter Carruthers

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 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. It 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 aren't 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 inrospectable *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 will 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 involving a form of introspection), whereas the fourth (which will be defended here) denies this. The present section will provide a brief explanation of each, before making some further introductory comments.

Model #1: two mechanisms. One possibility is that mindreading and metacognition are two independent capacities, realized in distinct cognitive mechanisms. This view has been elaborated and defended by Nichols and Stich (2003). 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 utilized 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 utilized 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 autistic people fit the second, thus confirming their account. These arguments will be discussed and evaluated in due course.

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, utilizing 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 considered above. This is that 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 where self-knowledge is compromised (because the instrospective 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 autistic people 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, will be examined in due course.

Model #3: metacognition is prior. A third sort of 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. Accounts of this kind have been proposed and defended by Goldman (1993, 2006), among others. They also lie behind much of the excitement surrounding the discovery of so-called "mirror neurons" (Gallese et al., 1996; Rizzolatti et al., 1996; Gallese and Goldman, 1998). 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, 1996b; Nichols and Stich, 2003), and he now maintains that introspection utilizes 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.

Model #3 makes predictions similar to those of model #2, but with an opposite valence. Both accounts agree that there should be cases where 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 that are utilized in mindreading are disrupted. Following Nichols and Stich (2003), Goldman (2006) argues that autistic people fit this profile.

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

[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 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.

versions of such an account have been proposed (Gopnik, 1993; Gazzaniga, 1995, 2000; Wegner, 2002; Wilson, 2002; Carruthers, 2006; some differences amongst 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, utilizing essentially the same inputs, which are all perceptual or quasi-perceptual in character (including visual imagery and "inner speech"—see Section 2 below). 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 described above endorses the existence of some or other variety of *introspection*, understood broadly to encompass any reliable method for forming beliefs about one's own mental states that *isn't* 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 isn't 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, will be one that accesses

[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 shall assume, as seems plausible, that basic capacities for both mindreading and metacognition are independent of our capacity for natural language.

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 Section 2 below). So also at stake in this target-article will be 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 above, 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 shall 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 above 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 Section 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 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; Dehaene and Naccache, 2001; Dehaene et al., 2001, 2003; Baars et al., 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 that 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 shall 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). Since 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 shall 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

[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.

that here, either. Indeed, I intend the discussion in this target-article to be neutral between proposed explanations of phenomenal consciousness.

While 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 will have 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 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. Since this event is apt to give rise immediately to a stored belief, it qualifies as a (perceptual) judgment. But since 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 seventeen 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, "Ten sixty-six, 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

[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.

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 that 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 co-evolved (Gomez, 1998; Origgi and 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 will need 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 of 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 the utterance’s linguistic properties (lexical meanings, syntax, and so forth) 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 aren’t ambiguous to us, in the way that the utterances of other 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 and Wilson, 1995). Since 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 insure 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), though 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 the section-title above 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, 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 while commissurotomy patients can often have 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 (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 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.

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 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 that 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 unwarranted, then. 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 (2008b). In consequence, this section will be 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

[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 to build and maintain a positive self-image. In fact, *both* accounts might be true.

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, since 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 and 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

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

both domains to emerge at about the same time. What this means is that developmental evidence is inherently incapable of 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 (Gopnik, 1993; Nichols and Stich, 2003; Goldman, 2006). 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 and 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 non-verbal looking-time and expectation measures. These show competence with false-belief understanding and other allegedly late-emerging aspects of mindreading capacity at around fifteen or twenty-four months, *long* before this had traditionally been thought possible (Onishi and Baillargeon, 2005; Bosco et al., 2006; Onishi et al., 2007; Southgate et al., 2007; Surian et al., 2007; Song and Baillargeon, forthcoming; Song et al., forthcoming). But no one has, as yet, been able to develop non-verbal 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 will show 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 and Meltzoff, 1997.) This will be some or other variant of the "Machiavellian intelligence" hypothesis (Byrne and Whiten, 1988, 1997; Dunbar, 2000), which points to the immense fitness advantages that can accrue to effective mindreaders amongst highly social creatures such as ourselves. And all should predict that one might expect to find simpler versions of mindreading capacity amongst other animals (perhaps confined to recognition of perceptual access and ignorance together with intention), especially amongst mammals who live in complex social groups. These predictions appear to be borne out (Hare et al., 2000, 2001; Tomasello et al., 2003a, 2003b; Cheney and Seyfarth, 2007; Hare, 2007; Call and Tomasello, 2008).

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 byproduct 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. Since neural connections are costly to build and maintain (Aiello and 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. In order 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.

All three of the competitor accounts need to tell some story about the evolution of introspection, then. What I shall 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 aren't 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 non-human animals (Terrace and Metcalfe, 2005). This topic will be 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—trouble-shooting and intervening in those processes in cases of difficulty, initiating new strategies, checking that tasks are proceeding as expected, and so on and so forth (Shallice, 1988). What I shall argue, however, is that while there is indeed a supervisory role for metacognition, it is one that doesn't require an introspective capacity distinct from the third-person mindreading system. I shall argue, in addition, that our metacognitive interventions aren't 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 and Perlis, 2005, for an especially egregious 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 (2005) 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 that we are discussing. But of course a very large body of genuinely metacognitive data remains, especially in the domain of metamemory (e.g. Nelson, 1992; Metcalfe and Shimamura, 1994). 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). And 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 equally well be done 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 sketched above.) 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.) While 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 and 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 and 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) aren't 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 non-human animals

The last few years have seen a flurry of experiments purporting to demonstrate the presence of metacognitive processes in non-human animals (Smith et al., 1995, 1997, 2003; Shields et al., 1997; Call and Carpenter, 2001; Hampton, 2001, 2005; Hampton et al., 2001; Smith, 2005; Son and Kornell, 2005; Beran et al., 2006; Washburn et al., 2006; Kornell et al., 2007). 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 (2008a), who demonstrates 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, 2008a). 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 (Festinger, 1957; Bem, 1967, 1972; Wicklund and Brehm, 1976; Nisbett and Wilson, 1977; Eagly and Chaiken, 1993; Wegner, 2002; Wilson, 2002). These data are consistent with a "dual method" account of metacognition (Goldman, 2006), according to which metacognition is 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 will be 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

[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.

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 will 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 that 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 Section 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 it 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 pre-motor cortex would also have such an effect; but it doesn't (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) there was actually no judgment made, or no 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 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 an hypnotic trance, and then later carry out an instruction given to them while hypnotized, will often confabulate an explanation for their action (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—e.g., "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—e.g., 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 shift 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 Festinger, 1957; Bem, 1967, 1972; Cooper and Duncan, 1971; Greenbaum et al., 1972; Wicklund and Brehm, 1976. For a more recent review, see Eagly and Chaiken, 1993.) Thus subjects who are manipulated into writing a counter-attitudinal essay for meager pay, but 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'm 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 and 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 isn't 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, while head-shaking is interpreted as disagreement with those reactions.

Now, it doesn't *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 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? This challenge will be taken up in the next section, where the only such account that I know of will be 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 that we have to our unconscious attitudes (whether or not they get expressed in speech or other imagery) is always interpretative, as argued above. 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 aren't epiphenomenal, but often make an important causal contribution to subsequent thought and behavior (Clark, 1998; Carruthers, 2002, 2006). 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 amongst those who work on the psychology of reasoning (Evans and Over, 1996; Sloman, 1996, 2002; Stanovich, 1999; Kahneman, 2002). 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 (Evans and Over, 1996; Frankish, 2004; Carruthers, 2009). 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 it is that the events that constitute System 2 achieve their characteristic effects. For only if they 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 shall 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,

[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.

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 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 that 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 that 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 doesn't 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 congressman." This utterance doesn't, 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 above. So in each case, then, although there is a conscious System 2 event to which I have introspective access, it *isn't* 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 that 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. Since 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 *aren't* accidental. The present section will evaluate 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 example, then 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.

While 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 prior to 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. Since 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 #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 aren't 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] For 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 that is 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

[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 that 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 and 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.)

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 (Wolpert and Kawato, 1998; Wolpert and Ghahramani, 2000; Grush, 2004). 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 described above 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 *isn't* involved (Jeannerod, 2006).

But why should the absence of a forward model lead subjects to feel that their actions aren't 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 (Weiskrantz et al., 1971; Blakemore et al., 1998). 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* to them that way.

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 aren't 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 aren't 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) aren't 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 (Hurlburt et al., 1994; Frith and Happé, 1999). 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 autistic people 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 autistic adults, 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 doesn't 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 weren't 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 autistic children 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 eleven 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 things 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).

The data from autistic people considered by Nichols and Stich (2003) and by Goldman (2006) don't support their introspectionist positions against an interpretative, "mindreading is prior", account, then. But there are other data that these authors don't discuss, which suggest that autistic people 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 autistic children 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 [e.g.] green one, or did you mean to hit the other [e.g.] 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 that 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 since the children *could* have answered the self-knowledge version of the

[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 amongst the easiest of mindreading tasks, one might predict that only very young or more severely disabled autistic individuals would be likely to fail at it.

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 autistic people provides 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 also will 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, 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 aren’t borne out by the data. In contrast, the central prediction made by 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. While 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.

Acknowledgements

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. They are: 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.

References

Aiello, L. and Wheeler, P. (1995). The expensive tissue hypothesis. *Current Anthropology*, 36, 199-221.

Anderson, J. (1995). *Learning and Memory: an integrated approach*. John Wiley.

Anderson, M. and Perlis, D. (2005). Logic, self-awareness and self-improvement: the metacognitive loop and the problem of brittleness. *Journal of Logic and Computation*, 15, 21-40.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B. (1997). *In the Theatre of Consciousness*. Oxford University Press.

Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science*, 6, 47-52.

Baars, B. (2003). How brain reveals mind: neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies*, 10, 100-114.

Baars, B., Ramsoy, T., and Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences*, 26, 671-675.

Baddeley, A., Chincotta, D., and Adlam, A. (2001). Working memory and the control of action: evidence from task switching. *Journal of Experimental Psychology: General*, 130, 641-657.

Barrett, L., Dunbar, R., and Lycett, J. (2002). *Human Evolutionary Psychology*. Princeton University Press.

Bayne, T. and Pacherie, E. (2007). Narrators and comparators: the architecture of agentive self-awareness. *Synthese*, 159, 475-491.

Begg, I., Duft, S., Lalonde, P., Melnick, R., and Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, 28, 610-632.

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

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

Benjamin, A. and Bjork, R. (1996). Retrieval fluency as a metacognitive index. In L. Reder (ed.), *Implicit Memory and Metacognition*, Erlbaum.

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

Birch, S. and Bloom, P. (2004). Understanding children's and adult's limitations in mental state reasoning. *Trends in Cognitive Science*, 8, 255-260.

Birch, S. and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18, 382-386.

Blakemore, S., Wolpert, D., and Frith, C. (1998). Central cancellation of self-produced tickle

sensation. *Nature Neuroscience*, 1, 635-640.

Block, N. (1995). A confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18, 227-247.

Bosco, F., Friedman, O., and 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.

Botterill, G. and Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge University Press.

Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L., and Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964-966.

Bratman, M. (1987). *Intentions, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. (1999). *Faces of Intention: selected essays on intention and agency*. Cambridge University Press.

Briñol, P. and Petty, R. (2003). Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology*, 84, 1123-1139.

Byrne, R. and Whiten, A., eds. (1988). *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.

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

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

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

Carruthers, P. (1996a). Autism as mind-blindness. In P. Carruthers and P. Smith (eds.), *Theories of Theories of Mind*. Cambridge University Press.

Carruthers, P. (1996b). Simulation and self-knowledge. In P. Carruthers and P. Smith (eds.), *Theories of Theories of Mind*. Cambridge University Press.

Carruthers, P. (2000). *Phenomenal Consciousness: a naturalistic theory*. Cambridge University Press.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25, 657-719.

Carruthers, P. (2006). *The Architecture of the Mind: massive modularity and the flexibility of thought*. Oxford University Press.

Carruthers, P. (2007). The illusion of conscious will. *Synthese*, 96, 197-213.

Carruthers, P. (2008a). Metacognition in animals: a skeptical look. *Mind and Language*, 23, 58-89.

Carruthers, P. (2008b). Cartesian epistemology: is the theory of the self-transparent mind innate? *Journal of Consciousness Studies*, 15.

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

Two Minds: Dual Processes and Beyond, Oxford University Press

Cheney, D. and Seyfarth, R. (2007). *Baboon Metaphysics: the evolution of a social mind*. University of Chicago Press.

Clark, A. (1998). Magic words: how language augments human computation. In P. Carruthers and J. Boucher (eds.), *Language and Thought*, Cambridge University Press.

Csibra, G. (2007). Action mirroring and action interpretation: an alternative account. In P. Haggard, Y. Rosetti, and M. Kawato (eds.), *Sensorimotor Foundations of Higher Cognition: attention and performance XXII*. Oxford University Press.

Damasio, A. (1994). *Descartes' Error: emotion, reason and the human brain*. Papermac.

Damasio, A. (2003). *Looking for Spinoza: joy, sorrow, and the feeling brain*. Harcourt.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1-37.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J., and Riviere, D. (2001). Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience*, 4, 752-758.

Dehaene, S., Sergent, C., and 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, 8520-8525.

Dennett, D. (1991). *Consciousness Explained*. Penguin.

Dennett, D. (2000). Making tools for thinking. In D. Sperber (ed.), *Metarepresentations*, Oxford University Press.

Descartes, R. (1637). *Discourse on the Method*. Many editions and translations now available.

Dunbar, R. (2000). On the origin of the human mind. In P. Carruthers and A. Chamberlain (eds.), *Evolution and the Human Mind*, Cambridge University Press.

Eagly, A. and Chaiken, S. (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich.

Edwards, G. (1965). Post-hypnotic amnesia and post-hypnotic effect. *British Journal of Psychiatry*, 111, 316-325.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.

Evans, J. and Over, D. (1996). *Rationality and Reasoning*. Psychology Press.

Farrant, A., Boucher, J., and Blades, M. (1999). Metamemory in children with autism. *Child Development*, 70, 107-131.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*, 44, 283-296.

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

Frith, C., Blakemore, S., and Wolpert, D. (2000a). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London*, B, 355, 1771-1788.

Frith, C., Blakemore, S., and Wolpert, D. (2000b). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Research Reviews*, 31, 357-363.

Frith, U. and Happé, F. (1999). Theory of mind and self-consciousness: what is it like to be

autistic? *Mind and Language*, 14, 1-22.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences*, 12, 493-501.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593-609.

Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, MIT Press.

Gazzaniga, M. (1998). *The Mind's Past*. California University Press.

Gazzaniga, M. (2000). Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition? *Brain*, 123, 1293-1326.

Gigerenzer, G., Todd, P., and the ABC Research Group. (1999). *Simple Heuristics that Make Us Smart*. Oxford University Press.

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, 16, 15-28.

Goldman, A. (2006). *Simulating Minds: the philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Gomez, J. (1998). Some thoughts about the evolution of LADS, with special reference to TOM and SAM. In P. Carruthers and J. Boucher (eds.), *Language and Thought*, Cambridge University Press.

Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1-14.

Gopnik, A. and 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 S. Parker, R. Mitchell, and M. Boccia (eds.), *Self-Awareness in Animals and Humans*, Cambridge University Press.

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

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1, 158-170.

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

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377-442.

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

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

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

Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences*, 1001, 134-144.

Hare, B. (2007). From nonhuman to human mind: what changed and why? *Current Directions in Psychological Science*, 16, 60-64

Hare, B., Call, J., Agnetta, B., and Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behavior*, 59, 771-785.

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

Harris, P. (2002a). What do children learn from testimony? In P. Carruthers, S. Stich, and M. Siegal (eds.), *The Cognitive Basis of Science*, Cambridge University Press.

Harris, P. (2002b). Checking our sources: the origins of trust in testimony. *Studies in History and Philosophy of Science*, 33, 315-333.

Hurlburt, R. (1990). *Sampling Normal and Schizophrenic Inner Experience*. Plenum Press.

Hurlburt, R. (1993). *Sampling Inner Experience with Disturbed Affect*. Plenum Press.

Hurlburt, R. and Akhter, S. (2008). Unsymbolized thinking. *Consciousness and Cognition*, 17, ??-??.

Hurlburt, R., Happé, F., and Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine*, 24, 385-395.

Jeannerod, M. (2006). *Motor Cognition*. Oxford University Press.

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

Kant, I. (1781). *The Critique of Pure Reason*. Many translations and editions now available.

Kazak, S., Collis, G., and 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.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609-639.

Koriat, A. (1997). Monitoring one’s own knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349-370.

Koriat, A., Ma’ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition. *Journal of Experimental Psychology: General*, 135, 36-69.

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

Kosslyn, S. (1994). *Image and Brain*. MIT Press.

Kreiman, G., Fried, I., and Koch, C. (2003). Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Science*, 99, 8378-8383.

Kruglanski, A., Alon, W., and Lewis, T. (1972). Retrospective misattribution and task enjoyment. *Journal of Experimental Social Psychology*, 8, 493-501.

Leslie, A. and Polizzi, P. (1998). Inhibitory processing in the false belief task: two conjectures. *Developmental Science*, 1, 247-253.

Levelt, W. (1989). *Speaking: from intention to articulation*. MIT Press.

Locke, J. (1690). *An Essay Concerning Human Understanding*. Many editions now available.

McCloskey, M. (1983). Naïve theories of motion. In D. Gentner and A. Stevens (eds.), *Mental Models*, Lawrence Erlbaum.

Metcalfe, J. and Shimamura, A. eds. (1994). *Metacognition: knowing about knowing*. MIT Press.

Nelson, T. ed. (1992). *Metacognition: core readings*. Allyn and Bacon.

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

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

Onishi, K. and Baillargeon, R. (2005). Do 15-month-olds understand false beliefs? *Science*, 5719, 255-258.

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

Origgi, G. and Sperber, D. (2000). Evolution, communication, and the proper function of language. In P. Carruthers and A. Chamberlain, (eds.), *The Evolution of the Human Mind*, Cambridge University Press.

Paulescu, E., Frith, D., and Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, 362, 342-345.

Phillips, W., Baron-Cohen, S., and Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, 16, 337-348.

Rey, G. (2008). (Even higher-order) intentionality without consciousness. *Review Internationale de Philosophie*, 62, 51-78.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L. (1996) Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, 131-141.

Russell, J. and Hill, E. (2001). Action-monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry*, 42, 317-328.

Schwartz, B. and Smith, S. (1997). The retrieval of related information influences tip-of-the-tongue states. *Journal of Memory and Language*, 36, 68-86.

Searle, J. (1992). *The Rediscovery of the Mind*. MIT Press.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge University Press.

Sheehan, P. and Orne, M. (1968). Some comments on the nature of post-hypnotic behavior. *Journal of Nervous and Mental Disease*, 146, 209-220.

Shergill, S., Brammer, M., Fukuda, R., Bullmore, E., Amaro, E., Murray, R., and McGuire, P. (2002). Modulation of activity in temporal cortex during generation of inner speech.

Human Brain Mapping, 16, 219-27.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge University Press.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.

Sloman, S. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, and D. Kahneman (eds.), *Heuristics and Biases: the psychology of intuitive judgment*. Cambridge University Press.

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

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

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

Smith, J., Shields, W., and Washburn, D. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26, 317-373.

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

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

Song, H. and Baillargeon, R. (forthcoming). Infants' reasoning about others' false perceptions. *Developmental Psychology*.

Song, H., Onishi, K., Baillargeon, R., and Fisher, C. (forthcoming). Can an agent's false belief be corrected through an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*.

Sperber, D. and Wilson, D. (1995). *Relevance: communication and cognition*. Second Edition. Blackwell.

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

Southgate, V. Gergely, G., and Csibra, G. (2008). Does the mirror neuron system and its impairment explain human imitation and autism? In J. Pineda (ed.), *The Role of Mirroring Processes in Social Cognition*, Humana Press.

Stanovich, K. (1999). *Who is Rational? Studies of individual differences in reasoning*. Lawrence Erlbaum.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite

measures to reveal them. *Acta Psychologica*, 106, 147-246.

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

Terrace, H. and Metcalfe, J. eds. (2005). *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*. Oxford University Press.

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

Tomasello, M., Call, J., and Hare, B. (2003b). Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, 7, 239-210.

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

Wegner, D. (2002). *The Illusion of Conscious Will*. MIT Press.

Wegner, D. and Wheatley, T. (1999). Apparent mental causation: sources of the experience of the will. *American Psychologist*, 54, 480-491.

Weiskrantz, L., Elliot, J., and Darlington, C. (1971). Preliminary observations of tickling oneself. *Nature*, 230, 598-599.

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

Wellman, H., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, 72, 655-584.

Wells, G. and Petty, R. (1980). The effects of overt head movements on persuasion: compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1, 219-230.

Wicklund, R. and Brehm, J. (1976). *Perspectives on Cognitive Dissonance*. Lawrence Erlbaum.

Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.

Wolpert, D. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3, 1212-1217.

Wolpert, D. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317-1329.