Chapter 26

Mindreading the self

Peter Carruthers

This chapter contrasts two different kinds of account of our knowledge of our own thoughts. According to standard theories, self-knowledge of at least a subset of thoughts is direct and non-interpretive. According to the alternative, which will be elaborated and defended here, self-knowledge results from turning our mindreading capacities on ourselves, relying on the same sensory channels that we employ for other-knowledge and utilizing many of the same sensory cues.

Introduction

Philosophers have traditionally assumed that self-knowledge is special. Knowledge of one's own thoughts, in particular (one's beliefs, judgments, desires, hopes, fears, decisions, and intentions) is supposed to be especially intimate, direct, and reliable. Indeed, Descartes (1641) famously believed that one's knowledge of one's own current thoughts is **infallible** (one cannot be mistaken about them), and that those thoughts themselves are **self-presenting** (to have them is to have infallible knowledge of them). Nor was Descartes by any means alone in holding such beliefs. Similar views were endorsed by Aristotle (see Caston, 2002), Augustine (see Bolyard, 2009; Mendelson, 2009), Locke (1690), and many others. While philosophers today don't endorse anything so extreme, almost all hold that knowledge of at least a subset of one's own thoughts is **authoritative** (incapable of being challenged by others) and **privileged** (arrived at in a special way that isn't available to others). Indeed, similar views are common even among cognitive scientists, especially those who believe that third-person mindreading capacities are grounded in first-person awareness (Gallese & Goldman, 1998; Rizzolatti, Fogassi, L., & Gallese, 2001; Goldman, 2006; Meltzoff & Brooks, 2008). In fact, some sort of tacit commitment to the special nature of self-knowledge has a strong claim to be a human universal. For although no work has been done on the subject by anthropologists in small-scale societies, it seems that such views have been endorsed across time and place (whether tacitly or explicitly) whenever people have reflected and written on the question (Carruthers, 2011).

The present chapter will suggest that this widespread view is radically mistaken. Far from being special, self-knowledge results from turning our mindreading abilities on ourselves. The same mental faculty that evolved for reading the minds of others and negotiating the social world gets turned toward the self, issuing in knowledge of our own thoughts (although often, also, in false beliefs about them). On this view, the mindreading faculty is arranged as one of the consumer systems for "globally broadcast" attended perceptual information (in the sense of Baars, 1988, 1997), for of course mindreading would need to have access to such information in order to perform its primary function. Plainly, attributing thoughts to other people requires observations of their behavior and physical circumstances. Self-knowledge can then rely on anything that is accessible through these same sensory channels, including one's own behavior and context, but also one's own visual imagery, inner speech, felt affect, and other forms of sensory experience. For imagery utilizes

the same mechanisms as does perception, and is globally broadcast in the same manner (Kosslyn, 1994; Kosslyn & Thompson, 2003). So while knowledge of our own sensory states is direct, knowledge of our own thoughts is just as interpretive in nature as knowledge of the thoughts of others, and relies on many of the same kinds of sensory cue.

Our discussion of these contrasting accounts will proceed as follows. In "Confabulation and dual-method theories," evidence of confabulation in attributing thoughts to ourselves will be discussed, providing one major strand of support for the views just outlined. This section will also consider the defensive moves that are available to defenders of the special character of self-knowledge. Then in "The interpretive sensory-access account" the interpretive sensory-access account of self-knowledge will be elaborated in somewhat more detail. "Dissociation data" considers potential dissociation evidence from schizophrenia and autism. "Brain imaging evidence" discusses some of the brain-imaging evidence that bears on the issue.

Confabulation and dual-method theories

More than half a century of careful research in social psychology has produced voluminous evidence that people will often **confabulate** about their own current or very recently past thoughts. That is, they issue reports of their current or recent thoughts that are manifestly false, but seemingly without any awareness of the falsity of their claims. Moreover, in many cases these reports are just the ones that third-party observers with access to the same information would attribute to the subjects, suggesting that in these instances, at least, self-attributions of thoughts result from turning one's mindreading abilities on oneself. Philosophers wishing to defend anything resembling the traditional view of self-knowledge have been forced to embrace **dual-method** accounts as a result of this data (Nichols & Stich, 2003; Goldman, 2006). They claim that sometimes we turn our mindreading abilities on ourselves (often resulting in confabulation), but on other occasions we have access to our own thoughts that is direct and non-interpretive. The present section will sketch some examples of confabulation, before evaluating the dual-method response.

Confabulation data

A significant portion of the evidence has been collected by those working within the "self-perception" framework initiated by Bem (1967). For example, Wells & Petty (1980) found that nodding one's head, while listening to a message on a tape significantly increases people's expressed agreement with the message thereafter, while shaking one's head while listening significantly decreases agreement. (Subjects were told that they were testing how well the headphones stay on people's heads, and that the message was incidental to the purpose of the experiment.) It seems that subjects interpret their own behavior as agreement or disagreement with the message, and adjust their reports of their own degree of belief in the subject of the message accordingly.

Briñol & Petty (2003) replicated this result, and were able to demonstrate that it is not a consequence of priming or positive mood caused by the head movements. They varied the persuasiveness of the message, finding that when the message is persuasive the original result replicates, whereas when the message is *un*persuasive the opposite occurs: those nodding their heads agree with the message even less, while those shaking their heads agree with it more. The experimenters were able to show that subjects interpret their head movements as agreeing or disagreeing with their own internal reactions while listening to the message (such as saying to oneself, "What an idiot!"), and they were able to find no evidence of mood changes.

Briñol and Petty (2003) also conducted a separate experiment in which subjects had to write three statements about themselves that might impact their careers, writing either with their right or their left hands. They were then asked for their degree of confidence in the statements that they had written. Right-handers who wrote with their left hands expressed significantly lower confidence, presumably because the shaky writing is interpreted by the mindreading system as a sign of hesitancy and uncertainty. Indeed, third parties who looked only at the written statements and were asked about the degree of confidence of the writer showed exactly the same effect.

These data are consistent with a mixed account, according to which reports of one's thoughts can be **influenced** by mindreading while *also* depending on some privileged channel of information. But many other results in the literature cannot easily be interpreted in this way. For example, Wegner & Wheatley (1999) asked subjects to report on their intentions in experiments in which (they believed) they were jointly controlling the cursor on a computer screen with another subject (who was in fact a confederate of the experimenters). Immediately after each trial they were asked to record the extent to which they had intended the final position of the cursor on a 100-point scale ranging from 1 ("I allowed the stop to happen") to 100 ("I intended to make the stop"). In one condition the confederate was told to play no part in the movement of the cursor, in fact giving subjects complete control. On average people still rated their degree of intent at only 56, just above the mid-point. Presumably, they made the reasonable assumption that control would be shared and therefore anchored on the mid-point of the scale, only adjusting upwards slightly in conditions in which they in fact had complete control (perhaps being sensitive to the presence of less resistance on the computer-mouse than they had expected).

This is already a remarkable result. For we can assume that the subjects made a decision to stop just prior to the time when they did (since the confederate played no role). We can also assume that they would have been paying close attention to their states of intending, since they knew that they would need to report on them immediately thereafter. But if their own decisions were directly available to them, then one would predict that they should have had a powerful sense of causality in these circumstances. For we know that in general temporally-contiguous events give people a strong sense of causation (McCloskey, Colebatch, Potter, & Burke, 1983; Young, 1995). The absence of any such effect speaks powerfully against the idea of direct introspective access to intentions.

In other conditions the confederate was instructed via headphones in such a way as to bring the cursor to a halt next to a particular type of object depicted on the screen (such as a beach ball), in circumstances in which the subject would hear the name of that type of object (ostensibly as a distracter). When the word was heard many seconds in advance of the stop, subjects on average scored the degree to which they had intended the stop at 45, presumably because they were sensitive to some resistance in the movements of the mouse in conditions in which the confederate in fact had ultimate control. But when the word was heard just before the stop, they scored their degree of intent at over 60. Presumably, their mindreading systems interpreted the coincidence of stopping near the object that had just been named as evidence of an intention to stop at that point.

Let me finish this brief sampling of confabulation data with some discussion of the "dissonance" tradition in social psychology, from which hundreds of supporting references could be provided. In a typical type of experiment, subjects will be induced to write an essay arguing for a conclusion that is the contrary of what they believe. In one condition, subjects may be led to think that they have little choice about doing so (for example, the experimenter might emphasize that they have previously agreed to participate in the experiment). In the other condition, subjects are led to think that they have freely chosen to write the essay (perhaps by signing a consent form on top of the essay-sheet that reads, "I freely agree to participate in this experiment.")

The normal finding in such experiments is that subjects in the free-choice condition (and only in the free-choice condition) change their reported attitudes on the subject-matter of the essay. This happens, although there are typically no differences in the quality of the arguments produced in the two conditions. If subjects in the free-choice condition have previously been strongly opposed to a rise in university tuition costs, for example (either measured in an unrelated survey some weeks before the experiment, or by assumption, since almost all people in the subject pool have similar attitudes), then following the experiment they might express only weak opposition or perhaps even positive support for the proposed increase. Such effects are generally robust and highly significant, even on matters that the subjects rate as important to them, and the changes in reported attitude are often quite large.

We know that freely undertaken counter-attitudinal advocacy gives rise to negatively valenced states of arousal, which dissipate as soon as subjects express an attitude that is more consistent with their advocacy (Elliot & Devine, 1994). Indeed, even *pro*-attitudinal advocacy will give rise to changes in expressed attitude in circumstances where subjects are induced to believe that their honest advocacy will turn out to have bad consequences (Scher & Cooper, 1989). In circumstances where subjects are offered a variety of methods for making themselves feel better about what they have done (an attitude questionnaire, a question about their degree of responsibility, and a question about the importance of the topic), they will use whatever method is offered to them first (Simon, Greenberg, & Brehm, 1995; Gosling, Denizeau, & Oberlé, 2006). For example, if asked first about the importance of the question of tuition raises, they will say that it is of little importance (even though in questionnaires administered a few weeks previously they rated it as of high importance), thereafter going on to express an unchanged degree of opposition to the change and rating themselves as highly responsible for what they did.

The best explanation of these patterns of result is that subjects' mindreading systems automatically appraise them as having freely chosen to do something bad, resulting in negative affect. Then when confronted with the attitude questionnaire they rehearse various possible responses, responding affectively to each in the manner outlined byDamasio (1994), Gilbert & Wilson (2007), and others. They select the one that "feels right" in the circumstances, which is one that provides an appraisal of their actions as being significantly **less** bad. As a result of making that selection, their bad feelings go away. For example, saying (and hearing themselves say) that they do not oppose a raise in tuition (contrary to what they believe) enables their earlier actions to be appraised as *not bad*, and as a result they cease to **feel** bad. In contrast, it seems quite unlikely that subjects should really be changing their minds prior to selecting an answer on the questionnaire, with their novel belief then being available to be authoritatively reported. For we know for sure that they do not change their beliefs unless offered the chance to express them, and there is no plausible mechanism via which a question about one's beliefs should lead to the formation of a new belief in these circumstances (which can then be veridically reported).

Such results are deeply problematic for traditional accounts of self-knowledge. For one would think that a direct question about one's beliefs (e.g. about the goodness or badness of a tuition raise, or about the importance of the issue) would have the effect of activating the relevant belief from memory. There seems no reason why a judgment of this sort should remain unconscious or be otherwise inaccessible to the subject. However, if subjects had authoritative access to this activated belief, then it would be mysterious how they could at the same time express an inconsistent belief and make themselves feel better by doing so. For if they say one thing while being aware that they think something else, then they would be aware of themselves as lying. And that ought to make them feel worse, not better.

Dual method theories

Someone wishing to defend a traditional account of self-knowledge might acknowledge the soundness of the data on confabulation, while pointing out that they only show that people **sometimes** attribute thoughts to themselves on the basis of self-directed mindreading. Consistently with the data, one can maintain that sometimes people have direct access to their thoughts, whereas on other occasions they rely on self-directed mindreading. Views of just this sort are defended by Nichols & Stich (2003) and Goldman (2006). Plainly, more needs to be said. For an account that simply asserts that sometimes we rely on self-directed mindreading and sometimes on introspection makes no predictions about when confabulation might be expected to occur, and it therefore cannot explain the patterning of the confabulation data. What is needed is some specification of the circumstances in which each of the two methods will be employed.

Nichols & Stich (2003) draw a distinction between **detecting** one's mental states and **explaining** one's mental states (or one's behavior). Introspection can only do the former. This is because explanation presupposes causation, and yet the causal relations that obtain among our thoughts, and between our thoughts and our actions, surely cannot be introspected. In contrast, while mindreading cannot directly detect a mental state, explanation falls squarely within its domain. What Nichols and Stich propose, then, is that subjects will resort to self-directed mindreading whenever they are asked **why** they did something or **why** they think something. In such circumstances confabulation will occur whenever the cues available for mindreading are misleading ones. In contrast, when asked simply to report on a current or recent thought, they should be able to access it directly, and confabulation will *not* occur.

The distinction between detecting and explaining might well be capable of accommodating some of the data on confabulation. But it plainly cannot capture it all. In particular, it cannot account for any of the examples discussed above. For in the self-perception and dissonance studies subjects are just asked to say how strongly they believe something. No explanations are required. It therefore remains a mystery why subjects should opt to mindread themselves when (by hypothesis) their thoughts are directly available for report. It might be felt, however, that the dual-control studies of Wegner & Wheatley (1999) are different. For in this case subjects are asked to divide responsibility for an outcome between themselves and another agent, which requires a judgment of their respective causal contributions. Even so, it remains puzzling that people's judgments of causality should anchor so closely around the midpoint in the subject-controlled trials. For if their intentions were accessible to them through introspection, as traditional accounts suppose, then the temporal contiguity between these and the outcome should have given subjects a powerful sense of control.

Goldman (2006) does not address the problem of explaining the patterning in the confabulation data. But Goldman (2009) opts to say that introspection is employed for conscious thoughts, whereas self-directed mindreading is needed for unconscious ones. All instances of confabulation are therefore explained as occurring in circumstances where the relevant thoughts are not conscious. There are, though, two broad kinds of account of conscious thought, and each makes Goldman's reply problematic. One claims that conscious thoughts are thoughts that we know ourselves to possess, either in general, or in the right sort of direct non-interpretive way. The other claims that conscious thoughts are ones that are "globally broadcast" to a wide range of executive, affective, and inferential systems. Let us consider these in turn.

If conscious thoughts are ones that we know ourselves to possess in some manner or other (whether by introspection or via self-directed mindreading), then it will be of no help to appeal to

the conscious–unconscious distinction in explaining instances of confabulation. For if conscious thoughts can involve self-directed mindreading then we should expect confabulation to occur in these cases too. On the other hand, if conscious thoughts are ones that we know ourselves to possess directly and non-interpretively, then we are no closer to saying in what circumstances confabulation can be expected. For it is already agreed that confabulation results from self-interpretation. So to say that confabulation may be expected in connection with unconscious thoughts is just to say that self-interpretation may produce errors in cases where we rely on self-interpretation. This is, of course, circular.

The remaining option for Goldman is to say that we have direct access to globally broadcast thoughts, whereas we need to rely on self-directed mindreading for the remainder. One problem for this option is that there is little evidence that thoughts (judgments, decisions, and the rest) are ever globally broadcast in the way that sensory or sensory-involving states are. For all of the evidence that we have of global broadcasting in the brain pertains to sensory states (Baars, 1988, 1997, 2002, 2003; Dehaene & Naccache, 2001; Dehaene, Naccache, Cohen, Bihan, Mangin, Poline, et al., 2001, Dehaene, Sergent, & Changeux, 2003, Dehaene, Changeux, Naccache, Sackur, & Sergent 2006; Baars, Ramsoy, & Laureys, 2003; Kreiman, Fried, & Koch, 2003). Another problem is that the best-validated models that we have of working memory, which also seems to employ a global broadcasting architecture, assume that it always implicates the maintenance, rehearsal, and manipulation of sensory-involving representations, including visual imagery and inner speech (Baddeley, 2006; Müller and Knight, 2006; Postle, 2006; D'Esposito, 2007; Jonides, Lewis, Nee, Lustig, Berman, & Moore, 2008). Moreover, it would be problematic, in any case, for Goldman to explain why the subjects' real thoughts in the confabulation experiments sketched in "Confabulation data" should *not* have been globally broadcast, if such a thing is possible at all. For everyone agrees that attention is the main determinant of global broadcast and entry into working memory, and in the circumstances of those experiments one would expect subjects to be attending to their judgments or intentions, since they were either asked directly about them, or knew that they would need to give a report just a few moments later.

I conclude that dual-method theories cannot account for the full extent of the confabulation data. As a result, the only theory that does so successfully is one that claims that self-directed mindreading is the only access that any of us *ever* has to our own thoughts. This account will be elaborated in Section 3, before some additional evidence is considered under "Dissociation data" and "Brain imaging evidence."

The interpretive sensory-access account

According to the interpretive sensory-access theory sketched in Section 1 and developed in detail in Carruthers (2011), the mindreading system is arranged as one of the consumers of globally broadcast sensory-involving information in the brain. It evolved initially for other-directed social purposes, whether of a "Machiavellian" sort (Byrne & Whiten, 1988, 1997), or for purposes of cooperation and collaboration (Richerson & Boyd, 2005; Hrdy, 2009), or both. This requires it to have access to perceptual information about the world, although by default it would also have access to any form of globally broadcast representation (including the attended outputs of proprioception and other forms of bodily experience). As a result, the mindreading faculty will also have access to imagistic representations (whether visual, motor, or in inner speech or hearing), since these utilize the same mechanisms as perception and can be globally broadcast in the same way (Paulescu, Frith, & Frackowiak, 1993; Kosslyn, 1994; Shergill, Brammer, Fukuda, Bullmore, Amaro, Murray, et al., 2002; Kosslyn & Thompson, 2003). This means that attributions of sensory

states to oneself are comparatively direct and immediate, since such states are available to the mindreading faculty as input.

It is important to realize that the mindreading system will have access to more than just strictly sensory non-conceptual states. This is because conceptual information of varying degrees of abstractness is generally bound into the content of any given sensory state and broadcast along with it. Thus, Kosslyn (1994), for example, characterizes the early stages of visual processing as a continual "questioning" of non-conceptual visual input by conceptual systems, which seek a "best match" with their representations of what objects and events of the relevant kind should look like. When a match is found, it is bound into the content of the visual percept to be broadcast along with it for yet other conceptual systems to consume and draw inferences from. In this way, there can be a cascade of increasingly abstract concepts bound into any given perceptual state, as successive conceptual systems receive the products of earlier systems' work, and categorize the input accordingly (Barrett, 2005). As a result, one doesn't just see textured surfaces and shapes, one sees *a face*; and one doesn't just see **a** face, one sees *one's mother*; and so on. Likewise for hearing: one doesn't just hear a stream of phonemes, one hears someone *calling one's name*, for example.

The work of the mindreading faculty, too, can be bound into the contents of globally broadcast perception or imagery. As a result, one doesn't just see someone's arm moving in the direction of a transparent object, one sees her as *reaching for a drink*; and one doesn't just hear a stream of phonemes when someone talks, but one hears him as *wanting to know the way to the church*; and so on, and so forth. Likewise one's own outer or inner speech can be heard as *judging that the church is straight ahead*. In either case the only access that this gives one to an underlying attitude is interpretive in character, depending on the combined work of the mindreading and language faculties. Yet, of course, an item of inner speech is not **itself** an attitude of any sort. So the event of hearing oneself as judging that the church is straight ahead is not itself an event of judging anything. Rather, at best, it expresses or is caused by such a judgment.

On the interpretive sensory-access account, then, while one generally has direct (non-interpretive) knowledge of one's own sensory-involving states, the only access that one has to propositional attitudes of judging, deciding, intending, and so on (whether one's own or someone else's) is interpretive, mediated by some form of sensory or imagistic awareness. The interpretive sensory-access theory comports well with global broadcasting accounts of the architecture of human cognition, as well as with widely accepted theories of working memory. It is also directly supported by the extensive confabulation data discussed earlier, since self-attributions of mental states will be subject to just the same sorts of errors of interpretation as attributions of mental states to other people. In contrast, no form of direct-access theory of self-knowledge has any of these benefits.

The interpretive sensory-access account is also supported by a widespread agreement among psychologists who study human metacognition or "thinking about [one's own] thinking" (including judgments of learning, feelings of knowing, and confidence judgments). This is that metacognitive judgments are inferential and cue-based, relying on a variety of sensorily-accessible cues (Reder, 1987; Metcalfe, Schwartz, & Joaquim, 1993; Koriat, 1995, 1997; Dunlosky & Metcalfe, 2009). People rely on such things as feelings of familiarity, or the swiftness with which an answer comes to mind, when judging whether they know something, or when judging their degree of confidence. There is nothing here to suggest that they have direct access to their underlying states of mind. Yet these findings are, of course, just what the interpretive sensory-access theory would predict.

All the evidence considered so far is strongly supportive of the interpretive sensory-access account. But it remains to consider some other evidence that might be thought, on the contrary, to support the distinctive and separate character of self-knowledge.

Dissociation data

One way of showing that the interpretive sensory-access account is incorrect would be to demonstrate dissociations in one's competence to acquire self-knowledge and other-knowledge. The account predicts that these should not occur, since each form of knowledge is held to employ the same mindreading faculty utilizing the same sensory channels (albeit sometimes relying on different forms of evidence, such as inner speech or visual imagery in the case of self-knowledge). Just such claims of dissociation have been made by Nichols & Stich (2003), Goldman (2006), and Robbins (2009) in respect of either schizophrenia, autism, or both. The present section will discuss these syndromes in turn.

Schizophrenia

There is now extensive evidence of mindreading deficits in schizophrenia generally (see Brüne, 2005, and Sprong, Schothorst, Vos, Hox, & Van Engeland, 2007, for wide-ranging reviews of the existing literature). Indeed, even first-degree relatives of people with schizophrenia show mindreading deficits that are independent of age, education, and IQ (Janssen, Krabbendam, Jolles, & van Os, 2003). So one might wonder whether people with schizophrenia *also* show deficits in self-knowing. If they do not, as Robbins (2009) speculates, then this would present an anomaly for the interpretive sensory-access account.

A test of this hypothesis is provided by Koren, Seidman, Poyurovsky, Goldsmith, Viksman, Zichel, et al. (2004), Koren, Seidman, Goldsmith, & Harvey (2006), who used the Wisconsin Card Sorting Task (WCST) in conjunction with measures of metacognitive ability. Following each sorting of a card (and before receiving feedback), patients were asked to indicate their confidence in the correctness of their performance on a 100-point scale, after which they had to indicate whether they wanted that trial to count toward their final score (which would impact how much money they would win). Koren and colleagues looked especially for correlations between the various measures of performance and other measures that are known to be predictive of real-world competence and successful independent living. (Specifically, they used measures of insight into one's own illness and measures of competence to consent to treatment.) They found only small-to-moderate correlations between the basic WCST scores and the latter. However, the results from the measures of metacognitive ability correlated quite highly with the measures of successful real-world functioning. These findings have since been confirmed by Stratta, Daneluzzo, Riccardi, Bustini, & Rossi (2009). And in a separate experimental paradigm, Lysaker, Dimaggio, Carcione, Procacci, Buck, Davis, et al. (2010) found that measures of metacognitive self-awareness are a good predictor of successful work performance of people with schizophrenia over a 6-month period.

It would seem, then, that self-directed metacognitive abilities are inversely related to the severity of schizophrenic illness. This allows us to conclude that metacognitive abilities are generally damaged in people with schizophrenia; for the severity of their disease correlates with an increased inability to monitor their current mental lives and to choose adaptively as a result. This is just what would be predicted if both self-knowledge and other-knowledge utilize the same mindreading faculty, as the interpretive sensory-access theory suggests.

Nichols & Stich (2003) claim that a specific form of schizophrenia—namely, passivity schizophrenia—demonstrates a dissociation in the reverse direction. They think that these patients exhibit a failure of self-knowledge together with normal mindreading abilities. The first part of this claim has at least a superficial plausibility. For such people complain that their own actions aren't under their control. A patient might say, for example, "When I decide to comb my hair, it isn't me who controls the movement of my arm, but the FBI." Such patients are also apt to complain of

"hearing voices" (in reality their own self-generated inner speech), and they may believe that other people are inserting thoughts into their heads against their will.

There are two things wrong with Nichols and Stich's suggestion, however. One is that there is no reason to think that people with passivity schizophrenia have normal mindreading abilities. In part this criticism is motivated by the very strong association between schizophrenia and mindreading deficits generally, as discussed above. But it is also supported by an fMRI study conducted by Brüne, Lissek, Fuchs, Witthaus, Peters, Nicolas, et al. (2008), specifically with patients suffering from passivity kinds of schizophrenic illness. While these people succeeded on the simple mindreading tasks they were asked to complete, they employed a very different network of brain regions when doing so than do normal controls. This suggests that their mindreading *system* isn't normal, even if they are partly able to compensate in other ways.

In the second place, however, classic passivity symptoms are not best explained by the failure of a self-knowledge system. Rather they are better explained by the failure of one of the main components of the action-control system (Frith, Blakemore, & Wolpert, 2000a,b). This is a comparator mechanism that is hypothesized to receive a so-called "forward model" of the expected sensory consequences of movement, created from the "efference copy" of the motor instructions for that movement, comparing this with the afferent sensory feedback from the movement itself, and enabling one to make swift on-line corrections as the movement unfolds (Wolpert & Kawato, 1998; Wolpert & Ghahramani, 2000; Jeannerod, 2006). We know that this system is damaged in passivity forms of schizophrenia specifically. For patients with passivity symptoms are unable to make online corrections in their own movements in the absence of visual feedback (Frith, 1992). There is reason to think that systematic damage to the comparator system would give rise to experiences of the sort that might well issue in a sense of alien control, as I shall now explain.

One of the normal effects of the comparator system is to "damp down" conscious experience of any incoming perceptual information that matches the predictions of the forward model. This is because if everything is proceeding as expected then no attention needs to be paid to it. As a result, sensory experience of one's own movements is normally greatly attenuated. This is why it is impossible to tickle yourself (Blakemore, Frith, & Wolpert, 1998, 1999). It is also why someone unwrapping a candy at the theatre will barely hear the noise they are making, while those around them are greatly disturbed. It turns out, however, that patients with passivity forms of schizophrenia **can** tickle themselves, and their experiences of their own actions **are not** modulated by their motor intentions (Blakemore, Smith, Steel, Johnson, & Frith, 2000). Hence, they will experience their own movements with the same sort of sensory vividness as would be present if someone else were making their movements for them, and they will experience their own inner speech just as if another person were speaking. This is, of course, exactly what they report.

I conclude, therefore, that there is no reason to think that patients with schizophrenia (or specific forms of schizophrenia) demonstrate a dissociation between self-knowledge and other-knowledge. There is nothing, here, to challenge the interpretive sensory-access account.

Autism

Nichols and Stich (2003) and Goldman (2006) argue that autism represents a dissociation between mindreading (which is widely agreed to be damaged in this population) and self-awareness, which they claim remains intact. They place considerable reliance on a study by Farrant, Boucher, & Blades (1999), who tested children with autism (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. It should be emphasized, however, that almost all of the children with autism who participated in this study were sufficiently well advanced to be able to pass first-level false-belief tasks. So we should **predict** that they would have some understanding of their own minds, too, and that they should be capable of completing simple metacognitive tasks.

Moreover, none of the experimental tasks employed by Farrant and colleagues required subjects to attribute current or recently past thoughts to themselves. On the contrary, the tasks could be solved by anyone who possessed the requisite mental concepts who was also a smart behaviorist. For example, one experiment tested whether the children with autism were aware that it is easier to learn a small number of items than a larger number. Not surprisingly, the children did well on this test. For 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 that are later evaluated as correct. (Note that the average age of the children with autism in this experiment was eleven years.)

In contrast with the claims of Nichols & Stich (2003) and Goldman (2006), many studies have found paired deficits of mindreading and self-knowledge among children with autism. Some of these have looked at children's awareness of their own intentions. Thus, Williams & Happé (2010) used the knee-jerk response, for example, asking groups of children whether or not they had **meant** to move their leg. The children with autism were much worse than the control groups in identifying their knee-jerk as unintended, and in all groups success was highly correlated with success in a set of third-person false-belief tasks.

In a separate set of experiments, Williams & Happé (2010) measured capacities to attribute intentions in the third-person as well as in the first. Subjects were asked to complete a picture, such as a drawing of a girl with a missing ear, or a cup with a missing handle. But in each case they drew on a sheet of transparent acetate that had been laid over another, so that although they **thought** they were completing one picture, they were in fact completing a different one. For example, in drawing what they intended to be the ear on the side of a girl's head they had in fact drawn a handle on a cup. When the ruse was revealed to them, they were asked what they had **meant** to draw. They then watched a video of the same task being undertaken by another child, and were asked the same question in the third person.

The results of this experiment were that the children with autism were significantly worse at identifying both their own and others' intentions than were the ability-matched children with developmental delay. In both groups success was strongly correlated with success in a number of false-belief tasks. It would appear from these data that the capacity to attribute intentions to oneself is just as damaged in children with autism as is the capacity to attribute intentions to other people, and that both result from the difficulties that such children have with mindreading in general.

Other studies have looked at the capacity to attribute false beliefs to oneself and to others, often using the unexpected contents test (or "Smarties task"). Typically-developing children begin to pass both versions of this task at about the same age, normally around four (Wellman, Cross, & Watson, 2001). A number of experimenters have found that children with autism are equivalently delayed on this task for both *self* and *other* (Baron-Cohen, 1991, 1992; Russell & Hill, 2001; Fisher, Happé, & Dunn, 2005). Some, however, have found that performance is significantly **better** on the *self* question than on the *other* question, suggesting that self-awareness might be comparatively spared in autism (Perner, Frith, Leslie, & Leekam, 1989; Leslie & Thaiss, 1992).

Williams & Happé (2009) reasoned that the differentially better performance on the *self* question found in some studies might be due to the fact that the children are asked at the outset to **say** what they think is in the container. Children with autism might then succeed in the task by remembering what they had previously said, rather than by recalling or reasoning about their earlier belief.

Williams and Happé therefore devised a version of the task that would elicit belief spontaneously, without requiring any verbal expression. The experimenter pretended at the outset of the interview to have cut her finger, and asked the subject to fetch her a band aid, in circumstances where a number of different types of container were in plain sight, but out of the experimenter's reach. When the child opened the band-aid box, however, he would find that it contained crayons. The same *self* and *other* questions were then asked as usual. The results were that children with autism performed poorly in both versions of this task relative to controls.

In fact, Williams & Happé (2009) found that the children with autism experienced significantly **more** difficulty in the *self* version of the task than when predicting what another person would think. A similar finding is reported by Lombardo, Barnes, Wheelwright, & Baron-Cohen (2007). Their subjects with autism had significantly more impairment in measures of understanding their own emotions than they displayed with regard to other people's emotions. These findings might be thought to suggest a partial dissociation between self-knowledge and other-knowledge. A more plausible suggestion, however, is made by Williams and Happé. This is that whatever rules and heuristics the children with autism have learned in order to help them cope, and to enable them to attribute mental states to people, will generally be outward-looking in character and focused on the social world. For it is the social world that they find especially threatening and unpredictable. So the difference may be one of performance, and does nothing to suggest that competence in mindreading can be spared relative to competence in self-attribution.

Finally, it is worth mentioning some studies by Klein, Chan, & Loftus (1999), Klein, Cosmides, Costabile, & Mei (2002), Klein, Cosmides, Murray, & Tooby (2004) of an individual with autism, which are claimed to demonstrate a dissociation between self-knowledge and other-knowledge. Although this individual has severely impaired episodic recall, and fails to distinguish among the personalities of close family members, he has a stable model of his own personality traits that correlates pretty well with the estimates of those who know him best. It seems, then, that not only can reliable self-knowledge of traits be obtained in the absence of episodic memory, but also that it is independent of any capacity to gain knowledge of the personality traits of other people.

Knowledge of one's personality traits is not the same as knowledge of one's current or recently past thoughts, of course, which is our focus in this chapter. However, it might be thought to imply it. Knowing that people are acting selfishly, or generously, or stubbornly seems to require knowledge of their goals, as well as an understanding of their construal of the situation (their beliefs). So if our conceptions of people's personalities are built up gradually from our evaluations of their actions as they occur, then such knowledge would seem to presuppose a capacity to attribute current thoughts to the agents in question. It is not obvious, however, that one's beliefs about people's personalities are always constructed in this way, especially when that person is oneself. Rather, one's self-conception may initially be constructed, in whole or in part, from the evaluations of others. If one's parent comments, "Don't be so stubborn," this might lead one to encode, "I am stubborn." And once one has formed a stable self-conception, this will be apt to influence one's behavior in a self-fulfilling manner. Conceiving oneself to be stubborn, one will be apt to act stubbornly; believing oneself to be generous, one will be more likely to do generous things; and so on.

Thus, if the individual studied by Klien and colleagues had formed his self-conception in such a manner, then the degree of correlation with other's personality assessments of him can be explained without needing to suppose that he has the capacity to attribute current thoughts to himself at all. And we can also explain why his judgments of the personality traits of his family are comparatively undifferentiated. For it seems likely that children have many fewer opportunities to observe other peoples' personality-relevant evaluations of close family members than they are aware of receiving themselves.

Even if we suppose that the individual with autism studied by Klein et al. (1999, 2004) had developed his self-conception on the basis of piecemeal evaluations of his own actions, however, the discrepancy between his trait-knowledge for *self* and familiar *others* can be explained by the interpretive sensory-access account. Simplifying somewhat, in order to judge that other people are acting generously one needs to attribute to them knowledge that someone needs help, combined with sufficient motivation to provide that help despite significant costs to themselves. This will require mindreading. But in order to judge that one is oneself acting generously it is far from clear that one needs to **attribute** to oneself knowledge that someone needs help. Rather, the first-order fact that someone **does** need help will suffice, thereby **reflecting** one's knowledge without requiring one to **metarepresent** one's state of knowledge. And in order to know that one is overcoming a significant cost to oneself while providing that help one can rely on subjectively experienced feelings of affective conflict. These are, of course, only accessible to the mindreading system in the first person (consistently with the interpretive sensory-access account of self-knowledge). The discrepancy between this individual's knowledge of his own personality traits and the traits of his family members may therefore result from a difference in **performance**, not reflecting any difference in **competence** in attributing mental states within the two domains.

I conclude, therefore, that there is no reason to think that people with autism demonstrate a dissociation between self-knowledge and other-knowledge, any more than people with schizophrenia do.

Brain imaging evidence

A widespread consensus has emerged concerning the network of brain regions that is specifically implicated in third-person mindreading. These include the medial prefrontal cortex, posterior cingulate cortex, superior temporal sulcus, and temporo-parietal junction (Frith & Frith, 2003; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe, 2009; Lombardo, Chakrabarti, Bullmore, Wheelwright, Sadek, Suckling, et al., 2010). The question for us is whether the same, or a distinct, brain network is implicated in self-knowledge. We first consider studies that have paired *self* and *other* mental-state attribution tasks, before examining studies of metacognition.

Self and other

There have been remarkably few studies that have directly targeted our question. There have, however, been numerous imaging experiments of knowledge of personality traits in oneself and others (e.g. Kelley, Macrae, Wyland, Caglar, Inati, & Heatherton, 2002; Kjaer, Nowak, & Lou, 2002; Lou, Luber, Crupain, Keenan, Nowak, & Kjaer, 2004; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004; Pfeifer, Lieberman, & Dapretto, 2007; but see Gillihan & Farah, 2005, for a powerful critique of the assumptions made by such studies). These are of little direct relevance for us, since no one thinks that a personality trait is the sort of thing that one can directly introspect. Even if the initial acquisition of trait-knowledge requires thought-attribution, the adults in these studies are likely to have well-established models of their own personality traits, in which case they can answer questions about themselves directly from memory without needing to reason at all (Klein & Lax, 2010). It is small wonder, then, that many studies find different patterns of activation in the two conditions—albeit with very little consistency across experiments.

In one of the very few studies to contrast third-person mindreading with attribution of current mental states to oneself, Ochsner, Knierim, Ludlow, Hanelin, Ramachandran, Glover, et al. (2004) scanned subjects while they viewed a series of photographs, in three separate conditions. In one, they had to judge their own emotional reaction to the image (pleasant, unpleasant, or neutral). In another, they had to judge the emotional reaction of a character depicted within the image

(pleasant, unpleasant, or neutral). And in the third base-line condition they had to judge whether the photograph had been taken indoors or outdoors. Many of the regions of the mindreading network were found to be active in common between the *self* and *other* conditions. These included medial prefrontal cortex, posterior cingulate, and the superior temporal sulcus.

However, *self* judgments activated medial prefrontal cortex to a greater extent than did *other* judgments. This effect is likely to result from the fact that medial prefrontal cortex seems to be active whenever one processes social information generally (Saxe & Powell, 2006), and because one would expect deeper and more elaborated processing in relation to the self (Gillihan & Farah, 2005). *Other* judgments, in contrast, distinctively activated an area of left lateral prefrontal cortex, which the experimenters interpret as an area implicated in maintaining and manipulating information about the external world. *Other* judgments also differentially activated an area of visual cortex, which the experimenters interpret as resulting from the greater attention paid to visual stimuli when judging the emotional state of another person. So there is nothing in these findings to suggest the existence of distinctive mechanisms for self-knowledge.

Most other studies that purport to contrast *self* and *other* mental-state attribution have failed to pair other-directed mindreading tasks with attributions of current mental states to oneself. For example, Saxe, Moran, Sholz, & Gabrieli (2006) claim to find areas of both overlap and non-overlap for *self* and *other*. However, the design of their study is an odd one. The *other* conditions are intended to test for false-belief reasoning. Subjects were scanned while reading either a false-belief story or a story involving a false photograph or map. As one might expect, the main elements of the mindreading network were active in this condition, including medial prefrontal cortex and the temporo-parietal junction bilaterally. In the *self* condition, in contrast, subjects read a series of trait adjectives, and either had to judge whether or not the adjective applied to themselves, or whether it was positive or negative. Since this task doesn't require one to attribute any current mental states to oneself, people will either answer from memory (using a stable self-model), or by mindreading and generalizing from items in episodic memory.

Likewise, Lombardo et al. (2010) conducted an extensive imaging study with a self–other design. In each case mentalizing judgments were contrasted with physical judgments. In the *self* condition, subjects had to use a four-point scale to answer questions like, "How likely are you to think that keeping a diary is important?" This was contrasted with physical questions like, "How likely are you to sneeze when a cat is nearby?" The *other* condition was identical, except that the questions all related to the Queen. (This study was conducted in the UK.) Note, however, that subjects weren't asked to make judgments about their current thoughts and attitudes. Rather, they were asked to estimate what their attitudes **would** be toward various suggested possibilities (such as keeping a diary). Since these questions might be ones that some subjects had never previously considered, they might have had to engage in the same sort of simulative reasoning process that they would use when trying to determine the likely attitudes of another person. Moreover, other subjects might have been able to answer the *self* questions directly from memory (for example, if they knew that they update a diary every day).

I conclude that while very few studies have contrasted the brain regions involved in mindreading with those that are active when one attributes a current or very recently past mental state to oneself, what evidence there is supports the interpretive sensory-access account.

Metacognition in the brain

Although many investigations of metacognition in the brain have failed to find activity in the mindreading network, this is likely to be an artifact of the experimental designs that have been

used. For instance, Maril, Simons, Weaver, & Schacter (2005) set out to differentiate between feelings of knowing and tip-of-the-tongue states. Since these are both metacognitive in nature, the interpretive sensory-access theory predicts that the contribution made by the mindreading system should be washed out when either one is subtracted from the other. Even when the brain activations involved in both of these kinds of feeling were combined together by the experimenters, they were contrasted with the combined "know" and "don't know" responses. Of course these, too, are equally metacognitive. Likewise in the studies by Reggev, Zuckerman, & Maril (2011), when episodic and semantic feelings of knowing were combined together they were contrasted with the brain activity involved in the "don't know" response. Since both sets of conditions involve metacognitive states, the interpretive sensory-access account predicts that activity should not be seen in the mindreading network.

Quite different results can be obtained when metacognitive judgments are contrasted with first-order ones. For example, Chua, Schacter, Rand-Giovannetti, & Sperling (2006) investigated the brain regions that are active when subjects make metacognitive confidence judgments. They contrasted judgments of confidence with first-order judgments of recognition. One form of differential activity was found in orbitofrontal cortex. While this lies outside the mindreading network, it nevertheless makes good sense. For this is one of the main brain regions where affective feelings are represented, and **judgments** of confidence are often grounded in **feelings** of confidence. But in addition, differential activity was found in posterior cingulate cortex and in regions of medial and lateral parietal cortex that include the temporo-parietal junction. Although the authors themselves don't notice the point, these are vital elements of the mindreading network, as we noted earlier.

In a later study, Chua, Schacter, & Sperling (2009) contrasted metamemory judgments with two distinct kinds of first-order judgment, one of which consisted of judgments of recognition, as before, but the other of which involved judgments of facial attractiveness (which was used as an additional control). The investigators found differential activity in a number of areas. These included posterior cingulate and areas of medial and lateral parietal cortex that contain the temporo-parietal junction. But in addition they found activity in medial prefrontal cortex, which is also generally thought to be part of the mindreading network—albeit a region whose functions may also be somewhat more general. Almost all components of the mindreading network were thereby found to be active.

These results provide further support for the interpretive sensory-access theory, while being correspondingly problematic for those who believe that self-awareness is direct and independent of mindreading.

Conclusion

This chapter has contrasted two views of knowledge of one's own thoughts. According to the first, self-knowledge of at least a subset of thoughts is direct, non-interpretive, and especially reliable. According to the second, self-knowledge results from turning our mindreading capacities on ourselves, utilizing sensory-involving cues (including visual imagery and inner speech as well as perceptions of our own behavior). These cues need to be interpreted, just as the mindreading system needs to interpret sensory input when attributing thoughts to other people. We have seen that the interpretive sensory-access account comports well with global broadcasting theories of the architecture of cognition, as well as with sensory-involving theories of working memory, and that it can explain the widespread data on confabulation for thoughts collected by social psychologists. In contrast, a direct-access account cannot explain this data. Moreover, there is no convincing evidence of dissociations between self-knowledge and other-knowledge in either

schizophrenia or autism, and nor do there appear to be different brain networks implicated in the two forms of knowledge. So the interpretive sensory-access theory is currently better supported by the evidence.

Acknowledgements

Some of the material in this chapter is drawn from Carruthers (2011), with permission of the author and Oxford University Press.

References

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

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

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

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

Baars, B., Ramsoy, T., & Laureys, S. (2003). Brain, consciousness, and the observing self. *Trends in Neurosciences* **26**:671–5.

Baddeley, A. (2006). *Working Memory, Thought, and Action*. Oxford: Oxford University Press.

Baron-Cohen, S. (1991). The development of theory of mind in autism: Deviance and delay. *Psychiatric Clinics of North America* **14**:33–51.

Baron-Cohen, S. (1992). Out of sight or out of mind: Another look at deception in autism. *Journal of Child Psychology and Psychiatry* **33**:1141–55.

Barrett, H. (2005). Enzymatic computation and cognitive modularity. *Mind and Language* **20**:259–87.

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

Blakemore, S-J., Frith, C., & Wolpert, D. (1999). Spatiotemporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience* **11**:551–9.

Blakemore, S-J., Smith, J., Steel, R., Johnson, E., & Frith, C. (2000). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychological Medicine* **30**:1131–9.

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

Bolyard, C. (2009). Medieval skepticism. In: E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford. edu/archives/spr2009/entries/skepticism-medieval>.

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

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

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

Byrne, R. & Whiten, A. (Eds) (1988). *Machiavellian Intelligence*. Oxford: Oxford University Press.

Byrne, R. & Whiten, A. (Eds) (1997). *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.

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

Caston, V. (2002). Aristotle on consciousness. *Mind* **111**:751–815.

Chua, E., Schacter, D., & Sperling, R. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience* **21**:1751–65.

Chua, E., Schacter, D., Rand-Giovannetti, E., & Sperling, R. (2006). Understanding metamemory: Neural correlates of the cognitive process and subjective level of confidence in recognition memory. *NeuroImage* **29**:1150–60.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B* **362**:761–72.

Damasio, A. (1994). *Descartes' Error*. London: Papermac.

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

Dehaene, S., Changeux, J-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences* **10**:204–11.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D., Mangin, J., Poline, J., & Riviere, D. (2001). Cerebral mechanisms of word priming and unconscious repetition masking. *Nature Neuroscience* **4**:752–8.

Dehaene, S., Sergent, C., & Changeux, J. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences* **100**:8520–5.

Descartes, R. (1641). *Meditations on First Philosophy*. In E. Anscombe & P. Geach (Eds & transl.), *Descartes Philosophical Writings*. London: Thomas Nelson & Sons (1954).

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. London: Sage Publications.

Elliot, A., & Devine, P. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* **67**:382–94.

Farrant, A., Boucher, J., & Blades, M. (1999). Metamemory in children with autism. *Child Development* **70**:107–31.

Fisher, N., Happé, F., & Dunn, J. (2005). The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *Journal of Child Psychology and Psychiatry* **46**:409–19.

Frith, C. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale: Erlbaum.

Frith, C., Blakemore, S-J., & Wolpert, D. (2000a). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews* **31**:357–63.

Frith, C., Blakemore, S-J., & Wolpert, D. (2000b). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B* **355**:1771–88.

Frith, U., & Frith, C. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **358**, 459–73.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* **2**:493–501.

Gilbert, D., & Wilson, T. (2007). Prospection: Experiencing the future. *Science* **317**:1351–4.

Gillihan, S., & Farah, M. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin* **131**:76–97.

Goldman, A. (2006). *Simulating Minds*. Oxford: Oxford University Press.

Goldman, A. (2009). Replies to the commentators. *Philosophical Studies* **144**:477–91.

Gosling, P., Denizeau, M., & Oberlé, D. (2006). Denial of responsibility: A new mode of dissonance reduction. *Journal of Personality and Social Psychology* **90**:722–33.

Hrdy, S. (2009). *Mothers and Others*. Cambridge: Harvard University Press.

Janssen, I., Krabbendam, L., Jolles, J., & van Os, J. (2003). Alterations in theory of mind in patients with schizophrenia and nonpsychotic relatives. *Acta Psychiatrica Scandinavic*a **108**:110–17.

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

Jonides, J., Lewis, R., Nee, D., Lustig, C., Berman, M., & Moore, K. (2008). The mind and brain of short-term memory. *Annual Review of Psychology* **59**:193–224.

Kelley, W., Macrae, C., Wyland, C., Caglar, S., Inati, S., & Heatherton, T. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience* **14**:785–94.

Kjaer, T., Nowak, M., & Lou, H. (2002). Reflective self-awareness and conscious states: PET evidence for a common midline parietofrontal core. *NeuroImage* **17**:1080–6.

Klein, S., Chan, R., & Loftus, J. (1999). Independence of episodic and semantic self–knowledge: The case from autism. *Social Cognition* **17**:413–36.

Klein, S., Cosmides, L., Costabile, K., & Mei, L. (2002). Is there something special about the self? A neuropsychological case study. *Journal of Research in Personality* **36**:490–506.

Klein, S., Cosmides, L., Murray, E., & Tooby, J. (2004). On the acquisition of knowledge about personality traits: Does learning about the self engage different mechanisms than learning about others? *Social Cognition* **22**:367–90.

Klein, S., & Lax, M. (2010). The unanticipated resilience of trait self-knowledge in the face of neural damage. *Memory* **18**:918–48.

Koren, D., Seidman, L., Goldsmith, M., & Harvey, P. (2006). Real-world cognitive—and metacognitive—dysfunction in schizophrenia: A new approach for measuring (and remediating) more "right stuff." *Schizophrenia Bulletin* **32**:310–26.

Koren, D., Seidman, L., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S., & Klein, E. (2004). The neuropsychological basis of insight in first-episode schizophrenia: A pilot metacognitive study. *Schizophrenia Research* **70**:195–202.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General* **124**:311–33.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General* **126**:349–70.

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

Kosslyn, S., & Thompson, W. (2003). When is early visual cortex activated during visual mental imagery. *Psychological Bulletin* **129**:723–46.

Kreiman, G., Fried, I., & Koch, C. (2003). Single neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences* **99**:8378–83.

Leslie, A., & Thaiss, L. (1992). Domain specificity in conceptual development: Evidence from autism. *Cognition* **43**:225–51.

Locke, J. (1690). *An Essay Concerning Human Understanding*, J. Yolton (Ed.). London: J. Dent & Sons (1965, two volumes). .

Lombardo, M., Barnes, J., Wheelwright, S., & Baron-Cohen, S. (2007). Self-referential cognition and empathy in autism. *PLoSOne* **9**:e833.

Lombardo, M., Chakrabarti, B., Bullmore, E., Wheelwright, S., Sadek, S., Suckling, J., MRC AIMS Consortium, & Baron-Cohen, S. (2010). Shared neural circuits for mentalizing about the self and others. *Journal of Cognitive Neuroscience* **22**:1623–35.

Lou, H., Luber, B., Crupain, M., Keenan, J., Nowak, M., & Kjaer, T. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences USA* **101**:6827–32.

Lysaker, P., Dimaggio, G., Carcione, A., Procacci, M., Buck, K., Davis, L., & Nicolo, G. (2010). Metacognition and schizophrenia: The capacity for self-reflectivity as a predictor for prospective assessments of work performance over six months. *Schizophrenia Research* **122**:124–30.

Macrae, C., Moran, J., Heatherton, T., Banfield, J., & Kelley, W. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex* **14**, 647–54.

Maril, A., Simons, J., Weaver, J., & Schacter, D. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *NeuroImage* **24**:1130–8.

McCloskey, D., Colebatch, J., Potter, E., & Burke, D. (1983). Judgments about onset of rapid voluntary movements in man. *Journal of Neurophysiology* **49**:851–63.

Meltzoff, A., & Brooks, R. (2008). Self-experience as a mechanism for learning about others. *Developmental Psychology* **44**:1257–65.

Mendelson, M. (2009). Saint Augustine. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: <http://plato.stanford.edu/archives/fall2009/entries/augustine/>.

Metcalfe, J., Schwartz, B., & Joaquim, S. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**:851–61.

Müller, N., & Knight, R. (2006). The functional neuroanatomy of working memory: contributions of human brain lesion studies. *Neuroscience* **139**:51–8.

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

Ochsner, K., Knierim, K., Ludlow, D., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience* **16**:1746–72.

Paulescu, E., Frith, D., & Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature* **362**:342–5.

Perner, J., Frith, U., Leslie, A., & Leekam, S. (1989). Explorations of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development* **60**: 689–700.

Pfeifer, J., Lieberman, M., & Dapretto, M. (2007). "I know you are but what am I?": Neural bases of self- and social knowledge retrieval in children and adults. *Journal of Cognitive Neuroscience* **19**:1323–37.

Postle, B. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience* **139**:23–38.

Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology* **19**:90–138.

Reggev, N., Zuckerman, M., & Maril, A. (2011). Are all judgments created equal? An fMRI study of semantic and episodic metamemory predictions. *Neuropsychologia* **49**:3036–45.

Richerson, P., & Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* **2**:661–70.

Robbins, P. (2009). Guilt by dissociation: Why mindreading may not be prior to metacognition after all. *Behavioral and Brain Sciences* **32**:159–60.

Russell, J., & Hill, E. (2001). Action-monitoring and intention reporting in children with autism. *Journal of Child Psychology and Psychiatry* **42**:317–28.

Saxe, R. (2009). Theory of mind (neural basis). In: W. Banks (Ed.), *Encyclopedia of Consciousness*, Vol. 2, 401–10 Cambridge: MIT Press.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage* **19**:1835–42.

Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science* **17**:692–9.

Saxe, R., Moran, J., Sholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Scan* **1**:229–34.

Scher, S., & Cooper, J. (1989). Motivational basis of dissonance: The singular role of behavioral consequences. *Journal of Personality and Social Psychology* **56**:899–906.

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

Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology* **68**:247–60.

Sprong, M., Schothorst, P., Vos, E., Hox, J., & Van Engeland, H. (2007). Theory of mind in schizophrenia: Meta-analysis. *British Journal of Psychiatry* **191**:5–13.

Stratta, P., Daneluzzo, E., Riccardi, I., Bustini, M., & Rossi, A. (2009). Metacognitive ability and social functioning are related in persons with schizophrenic disorder. *Schizophrenia Research* **108**:301–2.

Wegner, D., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of the will. *American Psychologist* **54**:480–91.

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* **72**:655–84.

Wells, G., & Petty, R. (1980). The effects of overt head movements on persuasion. *Basic and Applied Social Psychology* **1**:219–30.

Williams, D., & Happé, F. (2009). "What did I say?" vs. "What did I think?": Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders* **39**:865–73.

Williams, D., & Happé, F. (2010). Representing intentions in self and other: Studies of autism and typical development. *Developmental Science* **13**:307–19.

Wolpert, D., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience* **3**:1212–17.

Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks* **11**:1317–29.

Young, M. (1995). On the origin of personal causal theories. *Psychonomic Bulletin and Review* **2**:83–104.