The Case for Simulation Theory

Alvin I. Goldman
Karen Shanton

To appear in: A. Leslie and T. German, eds.,
Handbook of 'Theory of Mind'

1. What Is Simulation and What Is the Simulation *Theory*?

'Theory of Mind' is the cognitive capacity for attributing mental states to self and others. The task for cognitive science is to identify the cognitive resources, operations, and heuristics characteristically used in executing this capacity. Although the label 'theory of mind' includes the word 'theory', a central point of contention is whether anything like a scientific theory of mental states is used by ordinary people when making attributions. The present chapter provides evidence for an alternative approach: the simulation theory.

What is simulation, considered as a mental operation? As understood here, it is a process of re-enacting, or attempting to re-enact, other mental episodes. This operation can be used across a range of cognitive tasks. A re-enacted episode might be one undergone by the subject herself or by someone else in the past, present, or future (or a hypothetical scenario). Thus, it covers pre-enactment as well as re-enactment understood literally.

Seated in my living room on a wintry day, I might imagine myself instead watching the surf on some sandy beach. What I am trying to do is undergo a visual experience that matches (as closely as possible) a visual experience I would have if I really were on the beach. Vision science tells us that what transpires in visual cortex when undergoing visual imagery can, to a considerable extent, match what goes on during genuine vision (Kosslyn and Thompson, 2000). This is what we call a mental simulation. This is a case of intra-personal simulation: trying to re-enact an event in one's own mind. In using simulation to read others' minds, however, one would try to re-enact their mental states. That's just how mindreading characteristically takes place, according to simulation theory (ST).

However, ST makes some qualifications. It does not claim that simulation is always used in mindreading, merely that simulation is the default procedure. It is the most basic and spontaneous method. Another qualification concerns the scope of simulational mindreading. ST accounts for third-person and first-person past, present and hypothetical mindreading but it is an implausible picture of how people classify their own current conscious states, e.g., current feelings of soreness in the throat. First-person current attributions require separate treatment -- though they certainly fall within the scope of the subject. Some clarifications of ST are also in order. First, our version of ST does not say that simulation exhausts the act of mentalizing. Rather, simulation is part of a process (many processes, at any rate) of mindreading, but the final phase of a mindreading process is always the formation of a belief about a mental state. Second, early studies of ToM centered on the attribution of propositional-attitudes like belief and desire. More recent treatments consider the whole gamut of mental states, including somatic and affective states, as comprising the proper province for a theory of mentalizing. We shall follow suit.

2. Developmental Evidence Against TT and in Favor of ST

A landmark discovery in the study of mindreading was Wimmer and Perner's (1983) finding that 3-year-old children characteristically fail mindreading tests that require subjects to attribute false beliefs. Four-year-olds, by contrast, get these attributions right. What explains the one-year improvement in performance? An initially popular account meshes neatly with the theory theory (TT). It holds that younger children lack the concept of false belief; they don't think of beliefs as states that can be mistaken. This fits the TT approach, according to which unobservable constructs like belief are part of a science-like theory. If a 3-year-old's concept of belief makes no room for false belief, this must indicate that his belief-theory differs from that of adults and 4-year-olds. Thus, according to TT, errors on false-belief tasks are to be explained in terms of a conceptual deficit, a deficit that is overcome by suitable theory change. This story has been championed by Perner, Wellman, Gopnik, and others (Perner, 1991; Wellman, 1990; Gopnik, 1993; Gopnik & Wellman, 1991; Gopnik & Meltzoff, 1997).

Results on two types of false-belief tasks ostensibly support the conceptual change story. One is the unexpected transfer task, in which Maxi places his chocolate in a certain location and goes out to play. While he is away, Maxi's mother transfers the chocolate to another location. When Maxi returns, where will he expect the chocolate to be? Although an adult will answer, "In the old location," 3-year-olds indicate the new location (where it actually is), thereby failing to attribute a false belief to Maxi. Another false-belief task that 3-year-olds fail is the deceptive container task. Children know that a certain type of tube usually contains Smarties. A child shown such a tube and asked what's in it will say "Smarties". In this test, however, after the child responds in this fashion, the tube is opened and pencils are revealed instead. When asked soon afterwards what she thought was in the tube when first asked, the child replies "pencils". She fails to ascribe a false belief to her past self. Four-year-olds, by contrast, have little trouble with these types of tasks.

Theory theorists also formulate their account in terms of "rule-use." Scientific theories consist of laws or generalizations, propositions of the form "Events of type X are usually accompanied, or followed, by events of type Y". If children are little scientists who gradually learn and modify their theories, then it's the set of rules they know or don't know that dictates their ToM performances. Some evidence in support of this idea seemed to come from the finding that, whereas children aged about 3 years are accurate in reporting whether a person had seen an event, they are less accurate in saying whether the person knew about the event (Wimmer et al., 1988). Older children are better at this task. Theory theorists said that the difference lies in older children knowing a "see/know" rule that the younger children haven't yet learned.

The story of conceptual change has buckled, however, in the face of new findings. One pivotal finding is that children as young as 15 months demonstrate comprehension of false belief in a nonverbal task. Using a violation-of-expectation method, Onishi and Baillargeon (2005) found that 15-month-old infants were sensitive to whether an observed actor had or had not been exposed to relevant information and hence would or would not have a false belief. If even 15-month-old infants grasp the concept of false

belief, the case for a conceptual deficit about false belief in explaining 3-year-olds' errors on verbal tasks is totally undermined.

Mitchell and colleagues reported analogous doubts about the TT approach, now focusing on the rule-use version. In a series of studies in the 1990s (Robinson and Mitchell, 1995; Robinson et al., 1999), a high proportion of (early) 3-year-olds made correct judgments linking perceptual access with a consequent state of knowledge. This shows that 3-year-olds grasp the "see/know" link well before they can pass a (verbal) test of false belief. This implies that other variables must be partly -- if not wholly -- responsible for their poor performance. Mitchell et al. (2009) now add the argument that the conceptual change story is inconsistent with the fact that young children start out giving systematically incorrect judgments on standard (verbal) false-belief tasks. If children started life with no knowledge of any relevant rule, as TT traditionally claimed, their performance should be around chance level. In fact they exhibit a very systematic pattern of errors.

If knowledge or ignorance of rules does not determine performance on false-belief tasks, what does control such performance? Mitchell et al. (2009) pinpoint the roles of simulation and salience. An early study by Mitchell and Lacohee (1991) illustrates the importance of salience. Children confronting the Smarties tube task were allowed to select and "mail" a photo of the object they named when first asked what was in the tube. This tangible token helped many of them respond correctly when asked about their earlier belief, reversing the usual finding. Similar experimental permutations by Saltmarsh, Mitchell and colleagues also highlighted the role of salience (Saltmarsh & Mitchell, 1998; Saltmarsh et al., 1995). The role of salience may be subsumed under a more general theory of why 3-year-olds often fare poorly on false-belief tasks, namely, weak inhibitory control in children of that age (Carlson & Moses, 2001). This is a "performance-deficit" explanation of the phenomenon as contrasted with TT's "competence-deficit" explanation. Additional empirical support for this kind of account is provided by Leslie and colleagues (Friedman & Leslie, 2004; Leslie & Polizzi, 1998). Four-year-olds have substantial difficulty with an avoidance version of the false-belief task, in which the mindreading target wants to avoid rather than approach a particular box. Since four-year-olds in general understand the central concepts in this version of the task (false belief and avoidance desire), this finding can't be squared with competence-deficit accounts of mindreading development. Instead, Leslie and colleagues argue, it reflects a performance-deficit (specifically, weak inhibitory control). All of these findings constitute formidable challenges to TT, even in the developmental literature that had seemingly been hospitable to it. Now we turn to positive evidence for simulation from the same literature.

ST is distinguished from TT (in part) by its claim that third-person mindreading involves the projection of one's own states onto third-person targets. It is a reasonable prediction of ST, then, that this kind of projection also occurs during infancy, when mindreading develops. TT, at least in its child scientist variant, offers a different account of mindreading development. Infants start by making observations about behaviors and later draw inferences from these observations to unseen – and unseeable – states of their

targets. Is there anything in the developmental literature that can help us decide between these two stories? Is there reason to think that infants depend more heavily on self-experience than observation (or vice versa) when learning to mindread?

One experimental finding is that infants who have personal experience in producing certain goal-directed behaviors are more likely to attribute such goals to others than infants who lack such experience (Sommerville & Woodward, 2005; Sommerville et al., 2005; Woodward, 2009). A possible explanation of this finding is that producing goal-directed behaviors provides infants with extra opportunities to observe goal-directed behavior. It provides them with more (but not different) information than they would get from regular observation. This explanation has been ruled out, however, by other experiments. Repeated observation of goal-directed behavior does not affect infants' goal attributions (Woodward, 2009). By contrast, self-experience does influence infants' attributions, even when it doesn't provide them with additional opportunities for observation. For example, infants who have experience wearing a blindfold are significantly less likely to follow the gaze of a blindfolded adult than infants who lack such experience (Meltzoff & Brooks, 2008). This occurs even though they can't observe themselves wearing the blindfold.

These results suggest that infants' attributions of mental states to others are affected by providing them with special, first-person information. This information is different in kind from the information they get via observation. This conclusion is inconsistent with the child-scientist TT account of mindreading development. However, it coheres nicely with ST. ST claims that third-person mindreading involves projecting one's own states onto others. Therefore, it actually predicts that mindreading will employ special, first-person information.

Exactly how does ST think mindreading employs this information? How does engaging in goal-directed behavior or wearing a blindfold facilitate attribution of goals or perceptions? In the blindfold case, ST's explanation is quite straightforward. Experience with being blindfolded teaches the infant that, when eyes are covered with a cloth, vision is obstructed. When he later simulates the blindfolded adult, he simulates obstructed rather than unobstructed vision. ST offers a similar explanation in the goal-directed behavior case. Experience with engaging in goal-directed behavior teaches the infant that intentions and physical movements combine to achieve goals. Once he has learned to combine intentions and motions himself, he can re-enact their combination in others.

Now, there is another possible explanation of these findings: experience provides the infant with special, first-person information, which he uses as a premise in theories about others' mental states. For example, the infant learns that blindfolds obstruct vision and uses this to construct a theory about blindfolded others' perceptions. At the moment, the empirical evidence does not decide between the ST explanation and this alternative explanation. However, there is at least one reason to prefer the ST explanation. On the alternative explanation, infants extract information from action and use it in perception. This raises difficult questions about translation. How is the extracted information translated from the action code to a perception code? The ST explanation can avoid these

questions entirely. According to ST, information extracted from action is redeployed in (simulative) action. Therefore, it need not be translated into any other codes. (Woodward (2009) proposes something like this.)

There are two possible conclusions to be drawn from the above developmental evidence. One conclusion is that the evidence tells decisively against child scientist TT. TT characterizes infants as little scientists, first making observations, then drawing inferences about others' mental states. The evidence, on the other hand, depicts them as projecting from their own experiences and mental states to others' experiences and mental states. The second conclusion is that the evidence provisionally supports ST, but more experimental work must be done to establish this conclusively. At a minimum, ST is consistent with these experiments and, unlike alternative explanations, it does not encounter translation problems.

In recent work, Apperly (2008) recommends abandoning the ST-TT dichotomy in favor of other frameworks. This recommendation strikes us as hasty. First, Apperly does not show that the ST-TT dispute is not testable in principle but merely that it is difficult to test. Second, Apperly critiques individual studies purporting to settle the ST-TT dispute with the apparent intention of showing that no "crucial" experiment has resolved the issue. But crucial experiments of this sort are rare in any science. Greatest progress typically comes by combining insights from separate experiments, often using distinct methodologies. Section 4 illustrates this kind of progress in the TT-ST debate by combining evidence from neuroimaging and neuropsychology. Third, Apperly correctly points out that ST and TT can tell the same story about how inputs to mindreading mechanisms are generated. However, this doesn't show that we should give up the ST-TT distinction. Rather, it shows that the ST-TT dispute is a dispute about the mechanisms of mindreading, not the generation of inputs to mindreading. This is entirely consistent with our treatment of mindreading. As illustrated by Figure 1, inputs fed into the simulation heuristic could just as easily be fed into a theorizing heuristic.

3. Mindreading the Future and Past States of the Self

Deciding what to do commonly requires us to anticipate the future feelings we would have if this or that option were chosen. "How would I feel tomorrow if I ate that third piece of cake tonight?" "How refreshed would I feel if I now took a short break from work?" To answer such questions I must mindread my future states. How does one go about doing such mindreading? There are illuminating studies in social psychology of what we do in such tasks, studies that reveal some proneness toward error.

When weekend shoppers at a supermarket try to decide how much they will want to eat next week, they generally underestimate the extent of their future appetites if they have just eaten a big meal before shopping (Nisbett & Kanouse, 1969). Similarly, if people are asked to make a forced choice that will leave them either passing up a candy bar or passing up being told answers to certain questions, they generally underestimate

how much the second will frustrate them. Gilbert (2006) relates an experiment by Loewenstein et al. (1994/1998) that establishes this fact:

> [R]esearchers challenged some volunteers to answer five geography questions and told them that after they had taken their best guesses they would receive one of two rewards. Either they would learn the correct answers … or they would receive a candy bar but never learn the answers…. [P]eople preferred the candy bar before taking the quiz, but they preferred the answers after taking the quiz. In other words, taking the quiz made people so curious that they valued the answers more than a scrumptious candy bar. But do people know this will happen? When a new group of volunteers was asked to predict which reward they would choose before and after taking the quiz, these volunteers predicted that they would choose the candy bar in both cases. These volunteers … simply couldn't imagine that they would ever forsake a Snickers for a few dull facts about cities and rivers. (Gilbert, 2006: 115-116)

How can we explain these errors? Gilbert offers a clear-cut simulational answer. What participants try to do in these cases is imagine themselves in their own future shoes: how hungry will they feel, how curious will they feel, and so forth? They then predict their future feelings by seeing what their imagination produces now. In other words, current "prefeelings" are used to predict future feelings.

Reliance on prefeelings appears to be a tolerably reliable prediction method. How I would feel if I discovered my partner in bed with the mailman? It is probably a good heuristic to imagine (e.g., visualize) the encounter and see what emotions or feelings surface during the imaginative act. This will provide a good mini-sample of how my body and I really would react (rise in blood-pressure, pupil dilation, etc.). But there are many situations in which this heuristic is unreliable. What imagination generates on a given occasion depends not only on the scenario fed to the imagination, but on other current conditions of body and mind.

What interests us for present purposes is not the reliability or unreliability of the prefeeling heuristic for self-prediction but the mere fact that it is a simulation heuristic. Now, the reader might inquire: Is this really a simulation heuristic? And are the descriptions of the foregoing cases really accurate? Maybe people instead use some predictive rule rather than simulation, as TT might propose. Assuming that the descriptions are accurate, these uses of prefeelings are simulations because they are attempts to re-enact (that is, pre-enact) future mental states in order to predict them. Our definition of simulation doesn't require success so attempted re-enactment qualifies as simulation. What about the alternative descriptions of the cognitive processes used in terms of rule-use?

We do not mean to suggest that rules of thumb never come into play in mental forecasting. Often prefeelings are only one step in a more complex process. Gilbert et al. (2002) describe a three-step process consisting of, first, imagined future events ("mental proxies"), second, hedonic reactions to these imaginings, and third, "temporal

corrections" that adjust for the target events' temporal location. The third step might well use rules. However, it is likely that there is a simpler heuristic that omits the third step. This is a more primitive, default heuristic, which is a good candidate for being part of our early evolutionary heritage for planning. Recall that we described (our version of) ST as holding that simulation is the default method of mentalizing, not the exclusive method.

We might think that remembering our past mental states is a simpler proposition than predicting our future states. To remember what we believed, desired, felt, etc. in the past, nothing as involved as simulation is required. We can just pull memories of our past beliefs, desires and sensations from memory storage. A reason to doubt this simple account is that we're often wrong about our past mental states. More tellingly, we tend to be wrong in a very specific way: our memories of our past states line up with our current responses to the circumstances in which the past states occurred. This pattern has been observed for memories of emotions (Levine, 1997; Levine et al., 2001), pain (Eich et al., 1985), political attitudes (Goethals & Reckman, 1973), opinions of significant others (McFarland & Ross, 1987), test anxiety (Safer, Levine & Drapalski, 2002) and grief (Safer, Bonanno & Field, 2001).

How can we explain these findings? First, how can we explain the fact that we're often wrong about our memories? If we can be wrong about our memories, memory retrieval can't be simply pulling fully formed memories from storage. Instead, it must be a reconstructive process (Loftus, 1974). Second, how can we explain the particular types of memory errors we see? Why do memories of past events tend to line up with current responses? This type of error is analogous to a common bias in third-person mindreading called egocentric bias. As we'll see in section 5, third-person egocentric bias is best explained in terms of simulation. A similar explanation also seems plausible in the first-person case. Suppose that I use simulation to remember how I voted in the midterm elections in 2002. Suppose also that my position on a decisive issue has changed since 2002. To remember my voting decision accurately, I have to inhibit – or "quarantine" – my current position on the issue. If I fail completely to quarantine it, it will skew my reconstruction (memory) of my decision toward the decision I would currently make.

Further evidence that memory reconstruction is simulationist comes from reports of memory facilitation and inhibition effects. Re-assuming the body posture assumed during an experience facilitates remembering the experience whereas assuming an incongruent posture inhibits it. For example, you will remember a recent trip to the dentist more quickly if you are lying down than if you're standing up with your hands on your hips (Dijkstra et al., 2007). TT can't explain these effects. The position you are in when you deploy a theory should not affect the speed with which you generate a conclusion. ST, on the other hand, can explain them. If you are already in the position you were in when you had an experience, you will have a head start on re-enacting it (e.g. you won't have to mentally rotate yourself into the position). This will speed up (facilitate) your re-enactment of the experience. If you are in an incongruent posture, on the other hand, you will have to run through additional imaginative steps. This will slow down (inhibit) your re-enactment.

Egocentric biases and facilitation/inhibition effects give us some reason to think that memory retrieval is simulationist. Are there any other reasons? Recall that mental simulation is (attempted) re-enactment of mental processes. Is there direct reason to think that the processes by which we remember past mental states (memory processes) are re-enactments of the processes that produced those states (remembered processes)? The short answer here is yes. For one thing, memory retrieval is commonly accompanied by a feeling of conscious re-living, or autonoetic consciousness (Gardiner, 2001; Tulving, 2002). The most straightforward explanation of this phenomenological resemblance between memory and remembered processes is that memory retrieval involves (attempted) re-enactment of remembered events.

This explanation also accounts for neural resemblances between the two types of processes. Using single-cell recording (Gelbard-Sagiv et al., 2008) and fMRI (Cabeza et al., 2004; Sharot et al., 2004; Wheeler et al., 2000), researchers have shown that neural units and regions that are activated during experience of an event are selectively reactivated during retrieval of memories of the event. For example, one of Gelbard-Sagiv et al.'s participants displayed consistent activation in a single unit in the right entorhinal cortex when watching a clip from The Simpsons. The same unit was selectively reactivated when he remembered the clip. This suggests that memories are neural re-enactments of remembered events.

Evidence that memory is susceptible to egocentric biases and body posture facilitation / inhibition effects is evidence that it involves simulation. After all, ST has natural explanations for both of these phenomena while TT doesn't have an obvious story about either. This already compelling case is further bolstered by evidence of phenomenological and neural resemblances between memory and remembered processes. Such resemblances are exactly what we would expect to see if memory processes were re-enactments – or simulations – of remembered processes.

4. Third-Person Low-Level Mindreading

The previous section focused on self-attribution; here we return to other-attribution, the principal subject-matter of ToM. As we have noted, early stages in the study of ToM centered on belief-desire psychology, especially belief, as the testing-point for rival theories. Recent work has shifted a bit toward other types of mental states such as emotional states (fear, disgust, anger), and bodily feelings (pain, touch, itch). Arguably, many of these kinds of states lack propositional contents; but this does not render them unfit subjects for analysis under the 'ToM' heading.

In the 1990s a window opened onto a possible new route to mindreading. This occurred via the discovery of mirror neurons, a class of neurons first discovered in the premotor cortex of macaque monkeys, using single-cell recordings (Rizzolatti et al., 1996; Gallese et al., 1996). Mirror neurons are activated both when a monkey executes a specific goal-oriented action – for example, grasping an object or holding it -- and when the monkey observes another individual performing the same action. Premotor activation

can be considered the neural basis of an intention to perform a motor act. Since the same motor intention occurs in both performer and observer, neurons with this execution-observation matching property were dubbed "mirror neurons." The process by which mirroring is effected is called a "mirror process" or a "resonance process." Since an observer re-enacts the same motor intention as the observed performer undergoes, a mirror process satisfies our definition of a mental simulation process -- in this case, an interpersonal process. Of course, neither actor nor observer need be aware that mirroring is occurring; and at least in the case of the observer, the mirroring event is almost invariably unconscious. Moreover, although an observer's motor plan is transmitted to appropriate muscles, muscular activation is normally inhibited so that no overt imitation occurs.

Interpersonal mirroring in itself does not constitute mindreading; if a being has no concepts of mental states it cannot attribute mental states and cannot mindread. But if a creature possesses the requisite conceptual repertoire and uses its own (mirrored) motor-intentions as the basis for mindreading, it could attribute motor intentions to another individual quite accurately. It is in the nature of mirroring that an observer's own state resembles or matches (to a substantial degree) a state of the observed target. The existence of a mirror neuron system in humans has also been established, using various techniques. Here too the observation of actions elicits activations in brain areas that code for the same motor plans or movements (see Buccino et al., 2004; Gallese, Keysers, & Rizzolatti, 2004; Rizzolatti and Craighero, 2004, for reviews). Therefore, as Gallese and Goldman (1998) conjectured, mirror-based mindreading may well occur in humans.

Does empirical evidence support this conjecture? Clear-cut evidence for mirroring-based reading of motor intentions is, by our lights, still inconclusive. The most favorable experimental evidence of this sort was provided by Iacoboni et al. (2005), but it is open to alternative interpretations (see Goldman, 2008). However, potential evidence for mirror-based mindreading is not confined to the domain of motor mirroring. Many additional types of mirroring have been confirmed in addition to motor mirroring. The label "mirroring" is not always applied to these other cases, but the findings reveal observation/execution matching phenomena that are functionally similar to matching phenomena in motor systems. The additional domains of mirroring include emotions (e.g., disgust and fear) and sensations (e.g., tactile sensations and pain). To test for mirroring of disgust, Wicker et al. (2003) first scanned participants while they inhaled foul odors (among others) through a mask, and then scanned the same participants while they merely observed others inhale foul odors. The same brain areas were selectively activated in both conditions, i.e., the left anterior insula and right anterior cingulate cortex, areas known from animal studies to be involved in experiencing disgust.

The Wicker et al. study per se does not provide evidence for mirror-based mindreading. However, when evidence from neuropsychology is added to the story, the thesis is well supported (Goldman and Sripada, 2005). Patient NK suffered insula and basal ganglia damage. On a questionnaire concerning the experience of various emotions, NK scored significantly lower than controls for disgust but not for anger or fear. He was also significantly and selectively impaired in the ability to recognize – i.e.,

mindread -- disgust in others through either facial or auditory cues (Calder et al., 2000). Similarly, Adolphs et al. (2003) reported a patient, B, with extensive anterior insula damage who was selectively impaired at recognizing – i.e., mindreading -- disgust when shown dynamic displays of facial expressions. The simplest explanation of these patients' selective recognitional deficits is their inability to mirror disgust. The specificity of the mindreading deficit is crucial. These patients did not have mindreading deficits for other emotions, only for the unique emotion for which they had a re-enactment deficit. This is powerful evidence that normal mindreading of disgust through facial and other perceptual cues is based on a mirrored re-enactment of disgust, (for discussion, see Goldman and Sripada, 2005; Goldman, 2006, chap. 6). Apparently, when a normal person attentively observes a disgust-expressive face, she undergoes a disgust-like feeling, or correlated cortical events, that resemble those of the target. These mirrored events prompt an attribution of disgust to the observed target.

A similarly compelling finding is associated with pain re-enactment and attribution. It has been shown that, when a painful stimulus is received, there is a reduction of motor excitability of the muscles adjacent to the location of the painful stimulus – a sort of "freeze response." Using transcranial magnetic stimulation (TMS), Avenanti et al. (2006) found a corresponding decrease in motor excitability (motor evoked potentials, MEPs) during the mere observation of needles penetrating the flesh of the back of a hand of a human model. No such inhibition was found when participants were shown either a Q-tip gently moving over the same area of the hand or a needle penetrating a tomato. Thus, there was apparent re-enactment of the pain in the observer when -- and only when – the observed model would have undergone a painful experience. Was mindreading associated with these re-enactments? Yes. Participants rated the intensity and unpleasantness of the pain presumably felt by the model on a visual analogue scale, where 0 centimeters indicated 'no effect' and 10 centimeters indicated 'maximal effect imaginable'. The pain ascribed to the model during deep penetrations was evaluated as more intense and unpleasant than during pinpricks. Thus, there is ample reason to infer that participants' ratings of the model's pain – clear instances of mental-state ascriptions – were causally based on the mirrored pain experiences they themselves underwent.

Following Goldman (2006), mirror-based mindreading may be called low-level mindreading. Since mirroring is a species of simulation, the finding of mirror-based mindreading supports the ST approach to one type of third-person mindreading. What is meant by "low-level" as opposed to "high-level" mindreading? Goldman (2006) offers several indicators of low-levelness: unconsciousness, automaticity, primitiveness, and so forth, proposals that may not be entirely satisfactory (Vignemont, 2009). Perhaps a better criterion of demarcation between low-level and high-level is that low-level mindreading is stimulus driven whereas high-level mindreading is typically reconstructive and hence memory driven. As we shall see in section 5, high-level mindreading tends to involve the imagination, which retrieves and permutes contents from memory. This tends to be a slower, more effortful process. However, the boundary between low-level and high-level mindreading may not be a sharp one, partly because the exact nature and boundaries of the imagination are not well defined.

In this section we have seen that stimulus-driven processes like mirroring (which by definition is re-enactive and hence simulational) are causally responsible for some third-person mindreading. If an experience is tokened in a model and then "transmitted" to an observer via the model's facial expression, vocalization, or stimulus conditions, the observer will re-enact that experience (usually below the threshold of consciousness) and project -- i.e., attribute – it to the model.

5. Third-Person High-Level Mindreading

There is substantial evidence that third-person high-level mindreading closely follows the pattern we described for first-person future mindreading. In particular, it involves imaginative construction of scenarios thought to operate in the target. In a prototypical example, the first stage of the imaginative construction is creation of a set of initial states (in the self) antecedently thought to correspond to states of the target (but not the specific state the mindreader wishes to ascertain). This is "putting oneself in the other's shoes." The second stage consists of feeding these inputs into one of the mind's operating systems and letting it output a further state. Finally, the mindreader "reads" or detects that output state and projects it onto the target, i.e., attributes it to the target. This closely parallels Gilbert's account of how one would answer the hypothetical question, "How would you feel if you discovered your partner in bed with the mailman?" -- except the "target" there is the self in a hypothetical state. Now let us illustrate the pattern with a third-person example: predicting someone else's decision.

Suppose your friend Greta has been dating two different guys off and on, and likes each of them quite a lot, though each with qualifications. Now Greta decides it is time to get married: which boyfriend will it be? How would you predict her choice? You know many of the features she likes and dislikes in each of them, and her beliefs about their traits and patterns of interaction with her. To arrive at a prediction, then, you might put yourself in her shoes (i.e., imaginatively adopt Greta's likes, dislikes, and beliefs), feed them into your decision-making system, and let it output a choice. Once this choice is generated, you classify it and attribute it predictively to Greta. The last step may be called "projection." This entire scenario is diagrammed in Figure 1.

In Figure 1, all shapes represent either mental states or cognitive systems of the mindreader. Some states are "ordinary" or "genuine" states, that is, not states generated by imagination or pretense. These genuine states are depicted by unshaded shapes. In this diagram all the genuine states are beliefs, depicted by oval shapes. In addition to these genuine states, there are also simulated, pretend, or E-imagined states, which are depicted by shaded shapes. In this diagram, the simulated states include likings (represented by shaded squares), beliefs (represented by shaded ovals), and a choice or decision (represented by a shaded triangle). Finally, one cognitive system of the mindreader is also depicted: a decision-making system (represented by the donut). Scanning the figure from left to right, the left-most shape represents a (complex) genuine belief of the mindreader about Greta, a belief about her likes and knowledge vis-à-vis her boyfriends. In the next column to the right are three shaded shapes, representing the

mindreader's simulation of Greta's likes and beliefs (or knowledge) – all directed, as it were, "at" Greta as the target being simulated. They are the product of the mindreader's adopting Greta's perspective, or putting himself or herself in Greta's shoes (with respect to this range of states). Moving to the right, the diagram depicts these simulated states being fed as inputs into the mindreader's decision-making system (the donut). Next, the decision-making system outputs a choice, the content of which is "marry Ed." Since the mindreader is not "really" deciding to marry Ed – it's just a pretend, or simulated, choice made from Greta's perspective -- the triangle representing the decision is shaded. Finally, the mindreader detects this choice-state and ("really") attributes it to Greta, yielding the right-most belief: an attribution to Greta that she will decide to marry Ed.

This entire diagram is a simulation-theorist's hypothesis about what transpires in typical episodes of third-person high-level mindreading. Supplying evidence in support of this hypothesis is a separate task. In a first installment of such evidence we shall examine a pattern of errors or biases (egocentric biases) that characterize third-person high-level mindreading. Similar evidence was encountered for first-person future mindreading in section 3.

Keysar et al. (2003) had participants play a communication game in which a "director" instructed other players to move certain objects around a grid. The players first hid an object in a bag, such as a roll of tape. They – but not the director – knew what was in the bag; and they knew that the director didn't know. When the director told the other players, "move the tape," there were two candidate tapes he could have been referring to: a videotape that both director and players could see and a secret roll of tape in the bag. Which tape should be moved at the director's instruction? If the other players read the director's mental state correctly, his instruction would be unambiguous. Nonetheless, adult players behaved "egocentrically." Despite being given ample evidence of the director's ignorance, they misinterpreted what the director said in terms of their own knowledge rather than the director's (partial) ignorance.

This is readily explained in simulationist terms. The other players simply allowed their knowledge to creep into the input-set they used when trying to simulate the director. Under the simulation story, this is an unsurprising confusion. An optimal use of simulation requires a mindreader to preserve firm separation between two or more bodies of mental representations -- two or more ledgers, or "accounts," one might say. One body of mental representations is their own "genuine" states, and the other body (or bodies) of mental representations is "pretend" states associated with one or more targets. If ST is correct, these bodies of genuine and pretend states will contain some elements very similar to one another. (Pretend desires are similar to genuine desires; pretend beliefs are similar to genuine beliefs; etc.) It should therefore be easy to confuse them -- and there may be a tendency for genuine states to displace pretend ones. Thus, ST predicts confusions of just this sort, with "egocentric" consequences for their mindreading conclusions. By contrast, theory theory makes no comparable prediction (at least not without supplementary assumptions that are by no means straightforward). TT's portrait of mindreading does not call for the use of simulated states that are readily open to "invasion" by genuine states.

Another example of egocentric errors comes from an early paper by Camerer et al. (1989). Well-informed people were asked to predict corporate earnings forecasts by other, less informed, people. The forecasters knew that their targets had less knowledge than they did. Hence, to be accurate, their forecasts better not be "invaded" by their own proprietary knowledge. Nonetheless, the forecasters failed to suppress their prior knowledge completely: their predictions partly reflected their proprietary knowledge. Camerer et al. dubbed this phenomenon "the curse of knowledge." Birch and Bloom (2003) report similar findings in children, also applying the label "curse of knowledge."

The foregoing studies present behavioral evidence that supports the proposition that mindreading accuracy decreases when people fail to quarantine – or inhibit -- their own mental states. Next we present evidence about a specific brain area that seems to be responsible for this kind of inhibition or quarantining. A patient has been identified with a deficit in inhibiting his own states and who therefore has trouble making accurate mental-state attributions.

Samson et al. (2005) report the case of patient WBA, who suffered a lesion to the right inferior and middle frontal gyri extending into the right superior temporal gyrus. This lesion included a region that Vogeley et al. (2001) found to be responsible (in another patient) for "inhibiting one's own perspective" (i.e., one's own states). WBA displayed egocentric errors on numerous mindreading tasks, including attributions of belief, visual experience, desire, and emotion. These errors can be traced to damage to the region responsible for inhibiting self-perspective.

A different type of evidence favoring simulation over theory (or rule-use) in high-level mindreading emerges from general reflection rather than specific empirical findings. It involves reflection on the choice of propositional contents attributed in third-person mindreading (Goldman, 2006: 175-180). Propositional attitudes are mental states consisting of two components: an attitude type (e.g., believe, desire, hope) and a propositional content (expressible by a that-clause, such as "that Johnny will be on time"). When attributing a propositional attitude, a mindreader selects both an attitude type and a proposition. Here we focus on contents, more specifically, sub-propositional contents that occur in attitude ascriptions. These are the concepts that make up a proposition, concepts that are linguistically expressed by nouns, verbs, and so forth. Although mindreaders do not (voluntarily) "choose" the concepts they use in understanding others' mental states, they do, in effect, make such selections. Now, if mentalizers proceeded in a theorizing spirit, wouldn't they contemplate the possibility that other thinkers deploy different concepts than they themselves do? Especially when attributing attitudes to young children or pets, it would not be unreasonable to suppose that others' concepts differ from one's own. Yet our default procedure is to use words that express our own concepts rather than fashioning new concepts tailor-made for "alien" thinkers. People don't generate and compare alternative "translation manuals" of other people's talk, in the fashion Quine (1960) describes. It doesn't occur to them to represent others as thinking, for example, in terms of undetached rabbit parts rather than rabbits. In short, to permute a phrase used by Davidson (1984), our default procedure is

to think of others as same-thinkers (as ourselves). This, however, is not how genuine theorizers would proceed. Anthropologists in foreign lands anticipate the possibility of unfamiliar customs and modes of thought. If children are little scientists in their thinking about mental states, as theory-theorists allege, why is their default procedure so different from that of anthropologists? Why is it so thoroughly rooted in egocentric content attribution?

6. Other Domains and Conceptions of Simulation

The (interpersonal) simulation idea is often expressed in two other phrases: putting oneself in another's "shoes" and taking another's "perspective." Are these anything more than suggestive metaphors? Do we really take other people's perspective, in any literal sense? When we observe someone seeing a certain scene, does our mind replicate his perceptual state, or even attempt to do so? Could that possibly be true, even if one takes into account active, relatively high-level features of perceptual activity such as selective attention? Yes, as shown dramatically by Frischen et al. (2009).

Mirror-neuron research indicates that observing another person's action activates corresponding motor areas in the observer's brain and primes similar actions. But a key issue remains. To what extent are another's action-related intentions encoded and how does this influence one's own action? To achieve a simulation of another person's action that results in empathic understanding of them, the observer should be prone to represent the world from the other person's viewpoint. This allocentric (third-person) representation may be quite different from the egocentric, or body-centered, representation that an observer uses to guide her own actions. For example, witnessing another person's goal-directed movement should activate appropriate selective attention mechanisms in the observer, which are used in the same way as by the agent. In everyday actions such as picking up an item from a cluttered desk, one must select the target item from distracters. Selection is achieved by simultaneously inhibiting the processing of distracting stimuli, thereby reducing interference from competing but irrelevant information. Other studies have shown that irrelevant distracting objects closer to the starting position of the hand interfere more with a reaching response and are therefore associated with greater inhibition. Frischen et al. (2009) investigated whether witnessing another person's selective reaching movements leads the observer to activate similar selective attention processes as the agent is utilizing. If the observer simulates the observed agent's frame of reference, she should most strongly inhibit distractors that are most salient for the observed agent rather than those that are most salient according to her own frame of reference. This is exactly what Frischen et al. found.

The Frischen et al. study did not directly address the question of mental-state attribution. It is also unclear whether mindreading in this type of case would be considered low-level or high-level mindreading. On the one hand, it would involve the motor mirror-system. On the other hand, relatively high-level mechanisms like selective-attention mechanisms would also play a role. For all of these reasons, we discuss this study in the present section rather than an earlier section devoted to low-level or high-

level mindreading. In any case, the study is a dramatic demonstration of an "empathy" process (using the term loosely, to include non-affective interpersonal simulation), which could easily lead to third-person mindreading as a simple next step. The observer would simply ascribe to the agent some perceptual-attentive state that she herself undergoes.

We have already seen that mirror processes are the neural substrates for low-level mindreading. What about the neural processes for high-level simulation? Several researchers (Schacter and Addis, 2009; Buckner and Carroll, 2006) propose the existence of a "core" neural network that critically underlies both episodic memory and prospection (projecting oneself into the future). Schacter and Addis characterize the network as consisting of medial prefrontal and frontopolar cortex, medial temporal lobe, lateral temporal and temporopolar cortex, medial parietal cortex including posterior cingulate and retrosplenial cortex, and lateral cortex. Buckner and Carroll suggest that this core network also makes contributions to theory of mind, or mindreading. All applications of the network involve something like simulation insofar as they involve mental displacement of one's attention from the current situation to another situation either temporally and/or personally removed from the actual one. Involvement of this network in the mindreading domain is the least well established of its applications. But we include it here (under the "other domains" heading) because of its potential significance to other simulation-related phenomena.

Among other applications, simulation of future events may play an important role in mental health. Schacter, Addis, and Buckner (2008) review evidence that simulations play an important role in psychological well-being. Sharot et al. (2007) observed a correlation between optimism and future-event simulation. Their data showed that participants (1) felt that positive future events were closer in time than negative future events, (2) rated positive events in the future as more positive than positive events from the past, and (3) indicated that positive future events were more intensely "preexperienced" than negative future events. Moreover, these effects were strongest in the most optimistic subjects. Williams et al. (1996) reported that suicidally depressed patients have difficulty recalling specific memories of past events and also generating specific simulations of future events. Past and future events generated by depressed patients in response to cue words lacked detail and were "overgeneral" relative to those produced by nondepressed controls.

We turn now to a simulation thesis far more ambitious than either the one presented in the first five sections of this chapter or to the core network thesis presented above. Barsalou (1999, 2008, 2009) depicts simulation as a basic computational mechanism of the brain, with a wide array of applications. As he presents it, "simulation constitutes a central form of computation throughout diverse forms of cognition, where simulation is the re-enactment of perceptual, motor and introspective states acquired during experience with the world, body and mind" (Barsalou, 2009: 1281). The re-enactment process is said to have two principal phases: first, the storage in long-term memory of multi-modal states that arise from perception, and second, the partial re-enactment of these multi-modal states for later representational use. Associative neurons capture feature patterns and later reactivate these patterns (or parts thereof) in the absence

of bottom-up stimulation. When retrieving a memory of a bicycle, for example, associative neurons partially reactivate the visual state active during its earlier perception. Applied to concepts and categorization, Barsalou suggests that, after experiencing a category's instances, a distributed multi-modal system develops to represent the category as a whole. Pulvermuller (1999), Martin (2007), Pecher and Zwaan (2005), and others are cited as providing important evidence for the proposed cognitive architecture.

It goes without saying that the simulation theory of mindreading is not committed to a program as far-reaching as Barsalou's. However, to the extent that simulation is found to be a versatile type of cognitive process or heuristic, this is likely to redound to the plausibility of a simulation-based approach to mindreading. As always, however, the devil is in the details.

8. Mindreading and Introspection

This final section concerns the process(es) involved in attributing current states to the self. Why does this topic belong in a chapter on the simulation approach to mindreading? First, mental states, including concurrent states, are regularly ascribed to the self. No story of mindreading is complete without an account of how this task is executed. We do not claim that simulation participates in (current) self-attribution. Nonetheless, self-attribution is important to ST because it is embedded in simulation-based attribution to others. More precisely, something like self-attribution occurs in the final stages of a simulation heuristic.

Consider the right-most side of Figure 1. It depicts a mindreader as attributing a mental state to another person. How is the attributed mental state selected? That is, how does the mindreader arrive at a particular description or classification of the state? For example, how does she select an attitude type and a content for propositional-attitude ascription? Figure 1 does not depict this in detail, but it is supposed to convey the idea that a mindreader inspects or monitors her current state and somehow classifies it. Using this classification, the state is attributed to the target.

However, readers may object, if the state is attributed to the target, why do we speak of self-attribution? Very good, let us not say that the state is attributed to the self. Rather, the state that gets classified (for purposes of other-attribution) is a state of the self. In Figure 1 the state of deciding to marry Ed (arrived at in simulative mode) is classified and projected onto the target. However, this is a state of the mindreader, in the mindreader's own mind. One of the steps of the simulation routine is to monitor this state and detect, or determine, its properties. For this reason, reading one's own (current) states occupies a pivotal position in third-person mindreading.

Our account of current self-attribution prominently invokes introspection (see Goldman, 2006). A less off-putting label for such an account might be "inner sense" or "self-monitoring." Indeed, "self-monitoring" might be a better label for several reasons, including the fact that direct detection of one's own states may occur even when they are unconscious, whereas "introspection" is traditionally reserved for direct detection of

conscious states. Nonetheless, we shall continue to use the term "introspection" in our discussion, partly because of familiarity. It must be conceded, however, that theories of introspection have a controversial history in both philosophy and psychology. A big knock against introspection is its historical claim to infallibility. Subsequent developments have raised doubts about its reliability or accuracy. Even philosophers, a group that traditionally favored a "privileged access" approach to self-knowledge, no longer express confidence in its pristine accuracy. It should therefore be stressed that introspection's accuracy is not a central point of contention here. In endorsing a (heavily) introspectivist view of current self-mindreading, we make no ambitious claims for its accuracy, certainly not its infallibility. Our thesis is simply that introspection is intensively used in many tasks of classifying one's own mental states.

Psychologists commonly adhere to inferentialist, interpretationist, or confabulationist approaches to self-attribution rather than introspectivism. Consider three samples of work that appear to support this family of approaches. First, split-brain patients of Gazzaniga (1995) made choices based on information provided only to the right hemisphere. When asked why these choices were made, they offered explanations (using the left hemisphere, which controls the production of speech) that had nothing to do with the information visually displayed to the right hemisphere: transparent cases of confabulation. Gazzaniga (2000) therefore postulated a left-brain-based module called the "interpreter," whose job is to give explanations of one's behavior by creating plausible theories. Second, Nisbett and Wilson (1977) mounted a display table at a shopping mall that exhibited four identical pairs of pantyhose. Passersby were asked which pair they preferred and why. An analysis of their choices revealed a statistically significant position effect but, when asked whether their choices were affected by position, almost all participants denied it. Instead they offered a variety of other explanations, such as the selected pair's superior knit, sheerness, or what have you. Clearly, they lacked introspective access to the genuine causes of their behavior, said Nisbett and Wilson; they only speculated on the causes. Third, Gopnik (1993) argued that, because 3-year-olds make similar mistakes in attributing false beliefs to themselves as in attributing such beliefs to others, they must use the same method for self- as third-person attribution, namely, theory-driven inference.

What should be concluded from these studies about introspection's role in self-attribution? The split-brain example certainly suggests that when someone is expected to explain his or her action but lacks suitable knowledge, there will be felt pressure to invent an explanation. When the left hemisphere doesn't know why he performed a certain action (because the choice was made by the disconnected right hemisphere) it will turn to inference and speculation. This certainly shows that inference or confabulation is sometimes used for self-explanation – at least in cases of brain deficits. But this hardly shows that introspection is never or rarely used to answer "Why did you do it?" questions. Nor does it show that introspection isn't the standard, or default, method of answering them.

Similarly, the results of the Nisbett-Wilson experiment do not warrant the anti-introspectivism conclusion drawn by the experimenters and much of the psychological

community. Consider the supposedly false beliefs respondents had about their reasons for preferring a certain pair of pantyhose. Participants cited causes of their preferences different from the scientifically detected position-effect cause. Does this mean that they must have been in error? No. First, events can have multiple causes. The fact that there was a positional cause of a preference does not show that there weren't any other causes of it. Second, when they said things like, "I prefer(red) that pair because of its superior knit," it doesn't follow that the explanation was erroneous simply because all pairs of pantyhose were identical. What's of interest here is whether a respondent <u>believed</u> that the preferred pair had superior knit. It is distinctly possible that respondents <u>did</u> have beliefs like this, which played a causal role in shaping their product preferences. They may have <u>thought</u> the products differed in their types of knit even if it wasn't so. And introspection might have been used to determine that these beliefs were present.

But didn't Nisbett and Wilson's study also show that people <u>failed</u> to detect the real mental causes of the shoppers' product preferences, namely, the positions of the preferred products in the display? Perhaps, but what does this prove? It proves that some truths about their mental states went undetected by the respondents. Specifically, they neglected to realize that their preference was statistically related to the position of the products in the display. But would this conflict with any reasonable version of introspectivism? No. No reasonable version would claim that whenever introspection is applied to a target mental state, it delivers all truths concerning that state, never leaving the subject ignorant of any such truth. In particular, no reasonable version of introspectivism would hold that introspection delivers every <u>causal</u> truth about mental states to which it has access.

Finally, we turn to Gopnik's anti-introspectivist argument. The argument depends on two assumptions. First is the assumption that the mistakes 3-year-olds make in false-belief tasks are the result of poor theoretical inference. We have already provided evidence that this isn't correct. Second is the assumption that 3-year-olds use the same method in answering first-person false-belief questions as third-person false-belief questions. What evidence supports this assumption? The main evidence is allegedly parallel performance by children on a variety of first- and third-person mindreading tasks, in which they are said to commit the same kinds of errors. However, Nichols and Stich (2003, pp. 168-192) persuasively rebut this parallelism story, a rebuttal summarized by Goldman (2006, pp. 236-237).

What introspection-involving thesis for first-person mental state ascription would be reasonable? A sensible introspectivism would not claim that people have (direct) introspective access to their <u>past</u> mental states. A query to someone about a past mental state tends to trigger memory search that generates a current output. This current memory state can then be introspected. Thus, introspection plays a role in arriving at an answer, but the past state itself is not "directly" introspected.

Second, there are numerous different kinds of mental states, and only a careless introspectivism would claim introspective access to all of them. For example, there are stored or dormant mental states, on the one hand, and occurrent or activated ones, on the

other. Moderate introspectivism would hold that only activated states are introspectible, not stored states. For example, suppose that you want to determine whether you have a plan or intention to take your summer holiday in July. You cannot introspect the plan if it is lodged in memory. Rather, you must first retrieve or activate it and then introspect the activated state.

If introspection offers only a partial account of self-attribution, universal introspectivism is unpromising. Goldman (2006) therefore offers a dual-method theory. Under this theory, introspection is the default method of attribution for currently activated states. Inference or confabulation is a fallback method, called upon when introspection is inappropriate or unavailable. Goldman (2006) offers more details of how introspection works, and defends it as a viable operation that cognitive science should not dismiss. It can be considered a quasi-perceptual operation, not dissimilar to operations labeled "interoception" that are widely accepted in cognitive science.

Another semi-introspectivist theory of self-attribution is offered by Carruthers (2009, 2010). He is more adamant about the partial "elimination" of introspection than informative about its operation in the domains for which he accepts it. He rejects introspective access (or self-monitoring) for activated propositional attitudes like (non-perceptual) judgments and decisions, but agrees that we can introspect broadly perceptual events, including perceptions, imagery, and somatosensory states.

How would Carruthers handle the vacation-intention example? Surely the introspectibility of such an intention (once retrieved) is fully compelling, and surely it is a propositional attitude. Carruthers might take recourse in his claim that inner speech is imagery, and imagery is introspectible. He might further claim that every such activated intention uses inner speech. But there is no independent support for the contention that all instances of activated propositional attitudes involve imagery. So the blanket claim that non-perceptual attitude tokens are non-introspectible rests on dubious speculation about the perceptual or imagistic constitution of activated attitudes. Moreover, if this speculation is granted, it emerges that activated propositional attitudes are routinely introspectible. So, what is initially advertised as massive "elimination" of introspection (Carruthers's own term) does not amount to elimination after all.

Finally, we revisit the importance of introspection (or self-monitoring) to third-person mindreading. Not all simulation theorists buy into this thesis. Gordon, one of the originators of ST, rejects the notion that simulation-based attribution involves a "transference" of a mental state found in the self onto the target (Gordon, 1995). But how can some such cognitive step be avoided? All parties agree that third-person attribution involves assignment of mental states to a target. How does a mindreader select (in the case of propositional attitudes, for example) a specific type and content? According to ST, the selection is made by running a simulation, letting it generate a (mental) upshot, and then detecting or determining the nature or character of that upshot. According to the present proposal, determining the nature of this upshot (to be projected onto the target) is made by introspection or self-monitoring. Without such a proposal, an

account of the simulation heuristic would be seriously incomplete. With it, however, ST is an attractive – and empirically well-supported – approach.

References

Adolphs, R., Tranel, D. & Damasio, A. (2003). Dissociable neural systems for recognizing emotions. *Brain and Cognition* 52(1): 61-9.

Apperly, I.A. (2008). Beyond simulation-theory and theory-theory: Why social cognitive neuroscience should use its own concepts to study 'theory of mind.' *Cognition* 107: 266-83.

Avenanti, A., Minio-Paluello, I., Bufalari, I. & Aglioti, S. (2006). Stimulus-driven modulation of motor-evoked potentials during observation of others' pain. *NeuroImage* 32: 316-24.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences* 22: 577-660.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology* 59: 617-45.

Barsalou, L.W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London: Biological Sciences* 364: 1281-89.

Birch, S.A.J. & Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science* 14: 283-6.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C.A. & Rizzolatti, G. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI study. *Journal of Cognitive Neuroscience* 16: 114-26.

Buckner, R. & Carroll, D. (2006). Self-projection and the brain. *Trends in Cognitive Sciences* 11(2): 49-57.

Cabeza, R., Prince, S., Daselaar, S., Greenberg, D., Budde, M., Dolcos, F., LaBar, K. & Rubin, D. (2004). Brain activity during episodic retrieval of autobiographical and laboratory events: An fMRI study using a novel photo paradigm. *Journal of Cognitive Neuroscience* 16(9): 1583-94.

Calder, A.J., Keane, J., Manes, F., Antoun, N. & Young, A.W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3: 1077-8.

Camerer, C., Loewenstein, G. & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy* 97: 1232-54.

Carlson, S.M. & Moses, L.J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development* 72: 1032-53.

Carruthers, P. (2009). Simulation and the first-person. *Philosophical Studies* 114(3): 467-75.

Carruthers, P. (2010). Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research* 80: 76-111.

Davidson, D. (1984). On saying that. In *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Dijkstra, K., Kaschak, M. & Zwaan, R. (2007). Body posture facilitates retrieval of autobiographical memories. *Cognition* 102: 139–49.

Eich, E., Reeves, J., Jaeger, B. & Graff-Radford, S. (1985). Memory for pain: Relation between past and present pain intensity. *Pain* 23: 375-9.

Friedman, O. & Leslie, A.M. (2004). A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science* 28: 963-77.
Frischen, A., Loach, D. & Tipper, S.P. (2009). Seeing the world through another person's eyes: Simulating selective attention via action observation. *Cognition* 111: 212-8.
Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119: 593-609.
Gallese, V. & Goldman, A.I. (1998). Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 2: 493-501.
Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8: 396-403.
Gardiner, J. (2001). Episodic memory and autonoetic consciousness: A first-person approach. In A. Baddeley, M. Conway & J. Aggleton, eds., *Episodic Memory: New Directions in Research*. Oxford: Oxford University Press.
Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga, ed., *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.
Gazzaniga, M. (2000). Cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition? *Brain* 123: 1293-1326.
Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R. & Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science* 322(5898): 96-101.
Gilbert, D.T., Gill, M.J. & Wilson, T.D. (2002). The future is now: Temporal correction in affective forecasting. *Organizational Behavior and Human Decision Processes* 88(1): 430-44.
Gilbert, D.T. (2006). *Stumbling on Happiness*. New York: Alfred A. Knopf.
Goethals, G. & Reckman, R. (1973). The perception of consistency in attitudes. *Journal of Experimental Social Psychology* 9: 491-501.
Goldman, A.I. (2006). *Simulating Minds*. New York: Oxford University Press.
Goldman, A.I. (2008). Mirroring, mindreading, and simulation. In J. Pineda, ed., *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition*. New York: Humana Press.
Goldman, A.I. & Sripada, C.S. (2005). Simulationist models of face-based emotion recognition. *Cognition* 94: 193-213.
Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16: 1-14.
Gopnik, A. & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
Gopnik, A. & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language* 7: 145-171.
Gordon, R. (1995). Simulation without introspection or inference from me to you. In T. Stone and M. Davies, eds., *Mental Simulation*. Oxford: Blackwell.
Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C. & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3: 529-535.
Keysar, B., Lin, S. & Barr, D.J. (2003). Limits on theory of mind use in adults. *Cognition* 89: 25-41.

Kosslyn, S.M. & Thompson, W.L. (2000). Shared mechanisms in visual imagery and visual perception: Insights from cognitive neuroscience. In M.S. Gazzaniga, ed., *The New Cognitive Neurosciences, 2nd Ed*. Cambridge, MA: MIT Press.

Leslie, A. & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science* 1: 247-53.

Levine, L. (1997). Reconstructing memory for emotions. *Journal of Experimental Psychology: General* 126(2): 165-77.

Levine, L., Prohaska, V., Burgess, S., Rice, J. & Laulhere, T. (2001). Remembering past emotions: The role of current appraisals. *Cognition and Emotion* 15(4): 393-417.

Loewenstein, G.F., Prelec, D. & Shatto, C. (1998). Hot/cold intrapersonal empathy gaps and the under-prediction of curiosity. Unpublished manuscript, Carnegie-Mellon University. Cited in Loewenstein, G.F. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116: 75-98.*

Loftus, E. (1974). Reconstructing memory: The incredible eyewitness. *Psychology Today* 8(7): 116-9.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology* 58: 25-45.

McFarland, C. & Ross, M. (1987). The relation between current impressions and memories of self and dating partners. *Personality and Social Psychology Bulletin* 13(2): 228-38.

Meltzoff, A.N. & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology* 44(5): 1257-65.

Mitchell, P. & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition* 39: 107-27.

Mitchell, P., Currie, G. & Ziegler, F. (2009). Two routes to perspective: Simulation and rule-use as approaches to mentalizing. *British Journal of Developmental Psychology* 27: 513-43.

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

Nisbett, R.E. & Kanouse, D. (1969). Obesity, food deprivation and supermarket shopping behavior. *Journal of Personality and Social Psychology* 12: 289-94.

Nisbett, R. & Wilson, T. (1977). Telling more than we can know. *Psychological Review* 84: 231-59.

Onishi, K.H. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308: 255-8.

Pecher, D. & Zwaan, R.A., eds., (2005). *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge: Cambridge University Press.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences* 22: 253-79.

Quine, W.V.O. (1960). *Word and Object*. Cambridge, MA: MIT Press

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27: 169-92.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3: 131-41.

Robinson, E.J. & Mitchell, P. (1995). Masking of children's early understanding of the representational mind: Backwards explanation versus prediction. *Child Development* 66: 1022-39.

Robinson, E.J., Champion, H. & Mitchell, P. (1999). Children's ability to infer utterance veracity from speaker informedness. *Developmental Psychology* 35: 535-46.

Safer, M.A., Bonanno, G.A. & Field, N.P. (2001). It was never that bad: Biased recall of grief and long-term adjustment to the death of a spouse. *Memory* 9: 195-204.

Safer, M.A., Levine, L.J. & Drapalski, A. (2002). Distortion in memory for emotions: The contributions of personality and post-event knowledge. *Personality and Social Psychology Bulletin* 28: 1495-507.

Saltmarsh, R. & Mitchell, P. (1998). Young children's difficulty acknowledging false belief: Realism and deception. *Journal of Experimental Child Psychology* 69: 3-21.

Saltmarsh, R., Mitchell, P. & Robinson, E. (1995). Realism and children's early grasp of mental representation: Belief-based judgments in the state change task. *Cognition* 57: 297-325.

Samson, D., Apperly, I.A., Kathirgamanathan, U. & Humphreys, G.W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain* 128: 1102-11.

Schacter, D.L. & Addis, D.R. (2009). On the nature of medial temporal lobe contributions to the constructive simulation of future events. *Philosophical Transactions of the Royal Society, B* 364: 1245-1253.

Schacter, D.L, Addis, D.R., & Buckner, R.L. (2008). Episodic simulation of future events: Concepts, data, and applications. *The Year in Cognitive Neuroscience, Annals of the New York Academy of Sciences* 1124: 39-60.

Sharot, T., Delgado, M. & Phelps, E. (2004). How emotion enhances the feeling of remembering. *Nature Neuroscience* 7: 1376-80.

Sharot, T. Riccardi, M.A., Raio, C. M. & Phelps, E.A. (2007). Neural mechanisms mediating optimism bias. *Nature* 450: 102 -5.

Sommerville, J.A. & Woodward, A.L. (2005). Pulling out the intentional structure of human action: The relation between action production and processing in infancy. *Cognition* 95: 1-30.

Sommerville, J.A., Woodward, A.L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition* 96: B1-11.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology* 53: 1-25.

Vignemont, F. (2009). Drawing the boundary between low-level and high-level mindreading. *Philosophical Studies* 144(3): 457-66.

Vogeley, K., Bussfield, P., Newen, A., Herrmann, S., Happe., F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R. & Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage* 14: 170-81.

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

Wheeler, M., Peterson, S. & Buckner, R. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences USA (Psychology)* 97(20): 1125-29.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V. & Rizzolatti, G. (2003). Both of us disgusted in *my* insula: The common neural basis of seeing and feeling disgust. *Neuron* 40: 655-64.

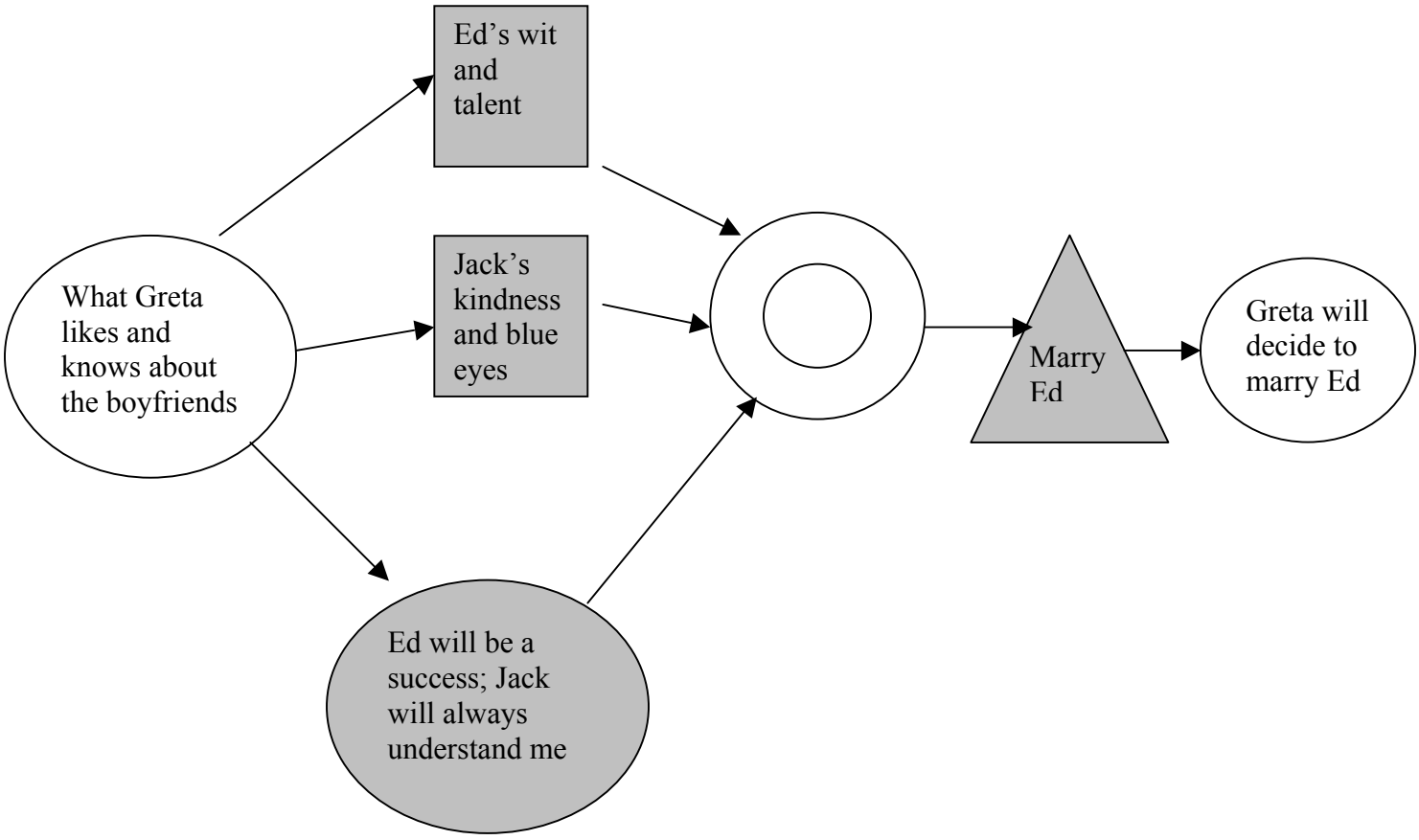
Williams, J.M.G., Ellis, N.C., Tyers, C., Healy, H., Rose, G. & MacLeod, A.K. (1996). The specificity of autobiographical memory and imageability of the future. *Memory & Cognition* 24: 116-25.

Wimmer, H., Hogrefe, G. & Perner, J. (1988). Children's understanding of informational access as a source of knowledge. *Child Development* 59: 386-96.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103-28.

Woodward, A.L. (2009). Infants' grasp of others' intentions. *Current Directions in Psychological Science* 18: 53-7.

Figure 1


Ed's wit and talent
What Greta likes and knows about the boyfriends
Jack's kindness and blue eyes
Marry Ed
Greta will decide to marry Ed
Ed will be a success; Jack will always understand me

APPENDIX

Can Unconscious States Be Introspected?

Alvin Goldman

The preceding discussion of introspection and first-person mindreading poses some interesting questions, especially when coupled with the earlier treatment of low-level third-person mindreading in section 4. Recall (manuscript, pp. 10-11) that different bodies of evidence concerning disgust and pain strongly suggest that when an observer attributes these mental states to a target (based on facial expressions, in one case, or painful stimuli applied to their bodies, in the other case) that the process of attribution includes the following two stages: (a) the observer undergoes a mirrored experience of a corresponding state in the target, and (b) the observer uses this self-experience to attribute the same state to the target. The evidence in the two cases is somewhat different, but the story inferred from the evidence is quite similar. In both cases self-experience is the basis for attribution. However, the hypothesized experiences in the mindreader – disgust in one case and pain in the other – are below the threshold of consciousness. So the mindreader appears to use his/her own unconscious mental state, appropriately categorized, to attribute a similar state to the other.

Now recall the story that the foregoing chapter tells about self-attribution of current mental states (based on the story developed in Goldman, Simulating Minds, 2006.) According to that story, the default procedure for self-attributing one's current mental states is introspection. Could the same introspective procedure also be used in the low-level mindreading of others, which makes use of mirrored states? In almost all cases an observer's mirror experiences occur below the threshold of consciousness.[i] But can unconscious states be introspected? Among researchers who endorse a process of introspection, almost all confine its operation to the conscious domain. Only conscious states are introspectible, not unconscious ones. This thesis, then, conflicts with our story of low-level third-person mindreading. So what further amendments -- if any -- are needed in our account of low-level third-person mindreading? Does it involve introspection, or a different operation? And if it involves another operation, how does the latter differ from introspection?

This is an interesting challenge: a challenge both to theories of introspection and to theories of consciousness. Let me first review salient features of the account of introspection offered in chapters 9-10 of Simulating Minds, to see whether they might be compatible with introspection of unconscious states. (That book makes only passing mention of the issue of whether introspection can be applied to unconscious states. It does not confront it squarely or extensively.)

Here are three central features of the proposed account of introspection. First, introspection is presented as a perceptual, or quasi-perceptual, process. It resembles perception insofar as it categorizes events or states in a "recognitional" rather than

inferential fashion. It acknowledges, as do most perceptual theorists, that introspection has no distinctive phenomenology of its own. But this holds equally of many perceptual processes, including internal perceptual processes widely accepted in cognitive science and referred to as "interoception".

Second, introspection is the core component of a cognitive system that uses a proprietary code to classify its targets (mental states). This is called the "introspective code". The idea is not dissimilar to Lycan's (1996) proposal, which regards introspective concepts as semantically primitive lexemes of the language of thought. In my case, the introspective code is said to represent both general mental categories, such as belief and intention, and also such qualities or magnitudes of mental states as location or intensity. Classifications of token mental states in one's own mind are outputs of the introspection operation.

Third, attention can be used to selectively direct or orient the introspective operation toward one or another mental state. This is analogous to what transpires in external perception when organs of perception (eyes, ears, etc.) are directed toward selected objects that bring them into suitable focus for enhanced processing and classification.

Do any of these features imputed to introspection exclude its applicability to unconscious states? I don't think so. Any of them can be understood (and should be understood) in a broadly cognitivist fashion that is neutral with respect to interacting with either conscious or nonconscious events in the brain. However, if pressed to identify the specific kinds of inputs to the introspective operation, the threat of inapplicability to unconscious states might well increase. Simulating Minds (pp. 246-253) considers four types of properties of mental states that might serve as inputs to the introspection system: (1) functional properties of mental states, (2) phenomenal properties of mental states, (3) representational properties of mental states, and (4) neural properties of mental states.

Indisputably, unconscious mental states have functional properties. But it is highly questionable whether an operation like introspection would have access – (relatively) "direct" access – to their functional properties. Functional properties are causal-relational properties. To detect such a property, introspection or another such monitoring process would have to be able to determine, of a targeted token state, which such causal-relational properties it has. That doesn't seem possible without a lot of access to the right sorts of relatum events, which introspection may well lack. Moreover, a good deal of inference would seem to be required, which falls outside the scope of activities usually imputed to introspection.[ii] To be sure, an operation like introspection might have access to the categorical bases of a state's functional properties. But this just kicks the can down the road. What is the nature of those categorical bases?

Perhaps the categorical bases are phenomenal properties, the second class of candidates. Even if their status as categorical bases of functional properties is dubious, they should still be considered as possible input properties. Here it seems straightforward that unconscious states would be excluded by this choice of inputs. On the standard

story, phenomenal properties, or qualia, are precisely what unconscious states lack. One could opt for the unorthodox view that unconscious states also have qualia, but that is a stretch. One loses a grip on what kinds of phenomena qualia are supposed to be when it is suggested that even unconscious states have them. In addition, even conscious propositional attitudes are often said to lack qualia. If so, there would be no relevant inputs to introspection in the case of the conscious attitude tokens. Yet introspection certainly classifies conscious attitudes and their properties along with other mental states (beliefs, strengths of belief, desires, preferences, intensities of preference, etc.).

Turn now to the third class of candidates: representational properties. One problem here, in my view, is that even if propositional attitude tokens possess representational properties (which accounts for their contents), they cannot be exhausted by representational properties. What makes something a believing or a desiring cannot consist in its being a (certain sort of) representation. Similarly for magnitudes of the attitudes such as strength or intensity; it is implausible that they are representations or representational. However, even if representational properties were assigned the status of inputs to introspection, this would not preclude introspection from being applied to unconscious states. Representational properties of unconscious states must be among the things that any broadly cognitive operation can interact with.

Finally, we turn to neural properties. This seems like a natural class of properties to serve as causal inputs to introspection. Obviously, their detectability is the same whether the states of which they are properties are conscious or unconscious. Which types of neural properties, of course, is a very large question, one that goes beyond the scope of the present set of reflections. If neural properties are the inputs to the introspective operation for conscious states, there is no bar to the same class of inputs being available for unconscious states. Neural properties are the category of choice in Simulating Minds. If we stick with this choice, the possibility of introspection being applied to unconscious states remains in play.

Thus far, then, there seems to be no bar to introspection being applied to unconscious states. Troubles may loom, however, from a different direction. If we tentatively agree to the thesis that unconscious states are introspectible, won't this run into trouble when we reflect more fully on what qualifies a state to be conscious or unconscious? Aren't there some theories of state consciousness, at any rate, that raise red flags?

In particular, consider the family of state-consciousness theories called "higher order" theories, including higher order thought (HOT) and higher order perception (HOP) theories. According to HOT and HOP, what makes a mental state a conscious state is its being the (intentional) object of a higher order state. In the case of HOP, it's a higher order (inner) perceptual state. In the case of HOT, it's a higher order reflective state. In either case it is assumed that the higher order reflection is an "unmediated" relation between the two states. Now, introspection is very naturally construed as precisely the type of relation that (non-mediately) links the higher-order state to the first-order state. Thus, if introspection can take unconscious states as its objects, then whenever this

occurs, a higher-order state will stand in the relevant sort of relation to the target unconscious state. But, then, according to the appropriate higher order theory, the unconscious state will be transformed into a conscious state. This would be unacceptable. Mirroring states that underlie low-level third-person mindreading are definitely (for the most part) unconscious. That is a non-negotiable datum. If any theory concerning such states threatens to turn them into conscious states, so much the worse for such a theory.

It seems, then, that we should try to avoid this outcome. An obvious way to do so, of course, is to reject any form of higher order theory of state consciousness. If being conscious or unconscious is a matter, say, of some intrinsic feature of a state, then a state's becoming an object of introspective mindreading simply won't affect its consciousness status. Similarly, if being conscious or unconscious is a matter of whether the state (or its content) is globally broadcast to other suitable sites in the brain, then, once again, its being an object of introspective mindreading won't affect its consciousness status. If I were more confident of the truth of either of these theories, I would rest more comfortably with these assurances.

Here's another way to escape the predicament: a stratagem for showing that it isn't possible to introspect unconscious states. According to the third feature of introspection that we advanced earlier, it is possible to attend to a state in order to introspect it. In Simulating Minds, this is called the inquiry facet of introspection (2006: 246). The second facet of introspection is the answering facet. This is the facet that takes relevant inputs and outputs a mental-state classification (or several of them). Now the inquiry facet of introspection is the directing of attention to selected mental states, and this, it may well be claimed, is a voluntary operation. At least sometimes it is a voluntary operation. Moreover, it is a voluntary operation of the conscious mind. This voluntary operation can only direct introspection to inspect conscious states, not unconscious ones. In the end, then, it seems that introspection can operate only in the sphere of the conscious. The unconscious plays its own game in a different ballpark.

Here are two possible replies. First, even if the inquiry facet of introspection only operates in the field of the conscious, this doesn't prove that the answering facet of introspection only operates in the field of the conscious. Maybe that facet of introspection operates (without any guidance from attention) in the field of the unconscious as well. Second, maybe what we ordinarily think of as attention is a particular mechanism (or family of mechanisms) that is restricted to the sphere of the conscious, but there is another mechanism of attention (or family of mechanisms) that operates in the sphere of the unconscious. And introspection can be guided by that mechanism too. This strikes us as a bizarre idea only because we aren't aware of the unconscious attentional mechanism.

Other possible approaches are certainly imaginable. For example, introspection might be restricted to the sphere of the conscious, but a different mechanism rather similar to introspection engages in unmediated mindreading of one's own states in the unconscious sphere. This might be called the dual-introspective-mechanisms approach.

This approach is rather unappealing. It multiplies mechanisms in an unconvincing fashion. In this respect it resembles a move (very briefly) contemplated by Nichols and Stich (2003: 160-161), the postulation of a separate monitoring mechanism for each separate type of mental state. As I argue in Simulating Minds (238-239), this would be an unpalatable profusion of mechanisms. Similarly, a duplication of introspection systems, one for the conscious realm and a second for the unconscious realm. seems unpalatable.

I have no final resolution to propose at the moment. I throw this out as a puzzle for anyone interested in mindreading and consciousness to contemplate. The evidence for low-level third-person mindreading looks very compelling. And it appears to require something like introspection as a sub-process. This is a troubling prospect when it is recognized that unconscious mental states would have to be introspected. Nonetheless, it's a major discovery of recent cognitive science that the unconscious mind executes a large number of tasks previously thought to be the exclusive preserve of the conscious mind. Introspecting one's own unconscious states may just be another activity that falls in this widening category.

References

Blakemore, Bristow, Bird, Firth, and Ward (2005). Somatosensory activation during the observation of touch and a case of vision-touch synaesthesia. Brain 128: 1571-1583.

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

Goldman, A. I. (2006). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press.

Keysers, Wicker, Gazzola, Anton, Fogassi, and Gallese (2004). A touching sight: SII-PV activation during the observation of touch. Neuron 42: 335-346.

Lycan, W. G. (1996). Consciousness and Experience. MIT Press.

Nichols, S. and Stich, S.P. (2003). Mindreading. Oxford University Press.

[i] In one exceptional case a patient was found who has "synaesthesia" for touch (Blakemore et al., 2005). When observing other people being touched, and therefore undergoing mirroring for touch (as is found in normal subjects; see Keysers et al., 2004), this patient consciously experiences the touch sensations that normal people would only undergo unconsciously. But this patient is highly unusual.

[ii] For discussion n of the kind of inference that would be required – and the threat of combinatorial explosion – see Goldman (1993).