Mirroring, Mindreading, and Simulation

Alvin I. Goldman

Department of Philosophy
Center for Cognitive Science
Rutgers, The State University of New Jersey
New Brunswick/Piscataway, NJ

To Appear In: Jaime Pineda, ed.,
Mirror Neuron Systems: The Role of Mirroring Processes In Social Cognition

Mirroring, Mindreading, and Simulation

Alvin I. Goldman

Abstract

What is the connection between mirror processes and mindreading? The paper begins with definitions of mindreading and of mirroring processes. It then advances four theses: (T1) mirroring processes in themselves do not constitute mindreading; (T2) some types of mindreading ("low-level" mindreading) are based on mirroring processes; (T3) not all types of mindreading are based on mirroring ("high-level" mindreading); and (T4) simulation-based mindreading includes but is broader than mirroring-based mindreading. Evidence for the causal role of mirroring in mindreading is drawn from intention attribution, emotion attribution, and pain attribution. Arguments for the limits of mirroring-based mindreading are drawn from neuroanatomy, from the lesser liability to error of mirror-based mindreading, from the role of imagination in some types of mindreading, and from the restricted range of mental states involved in mirroring. "High-level" simulational mindreading is based on enactment imagination, perspective shifts, or self-projection, which are found in activities like prospection and memory as well as theory of mind. The role of cortical midline structures in executing these activities is examined.

Keywords: Cortical midline structures; Emotion; Folk psychology; Intention; Mentalizing; Mindreading; Mirror neurons; Mirroring processes; Pain; Simulation; Self-projection; Social cognition; Theory of mind.

1. Introduction

Mirror systems are well established as a highly robust feature of the human brain (Rizzolatti, Fogassi and Gallese, 2004; Gallese, Keysers, and Rizzolatti, 2004; Iacoboni et al., 1999; Rizzolatti and Craighero, 2004). Mirror systems and mirroring processes are found in many domains, including action planning, sensation, and emotion (for reviews, see Keysers and Gazzola, 2006; Gallese, Keysers, and Rizzolatti, 2004; Goldman, 2006). Since mirroring commonly features an interpersonal matching or replication of a cognitive or mental event, it is a social interaction. It involves two people sharing the same mental-state type, although activations in observers are usually at a lower level than endogenous ones, commonly below the threshold of consciousness. There is strong evidence that mirror systems play pivotal roles in empathy and imitation (Iacoboni et al., 1999; Rizzolatti 2005; Iacoboni 2005; Gallese 2005; Decety and Chaminade 2005). Indeed, in a minimal sense of the term, 'empathy' might simply mean the occurrence of a mirroring process. In this paper, however, I shall focus on the connection between mirroring processes and another category of social cognition, viz., mindreading or mentalizing.

By 'mindreading' I mean the attribution of a mental state to self or other. In other words, to mindread is to form a judgment, belief, or representation that a designated person occupies or undergoes (in the past, present, or future) a specified mental state or experience. This judgment may or may not be verbally expressed. Clearly, not all judgments about other people are acts of mindreading. To judge that someone makes a certain facial expression, or performs a certain action, or utters a certain sound is not to engage in mindreading, because these aren't attributions of mental states. To attribute a mental state, the judgment must deploy a mental concept or category. Thus, if 'empathize' simply means 'echo the emotional state of another,' empathizing isn't sufficient for mindreading. A person who merely echoes another's emotional state may not represent the second person at all, and may not represent her as undergoing that emotional state (a species of mental state). It's certainly possible that mirroring processes are responsible for acts of mindreading as well as for imitation or empathy, but this involvement in mindreading doesn't "logically" follow from their role in either imitation or empathy. A connection between mirroring and mindreading must be considered separately.

2. Definitional Issues

The phrase 'mirroring process' can have either a wide sense or a narrow sense. In the wide sense, it refers to an interpersonal process that spans both sender and receiver. In a narrower sense, it refers to an intrapersonal process that includes only a receiver (a single individual). Unless otherwise specified, I shall understand 'mirroring process' in the narrow sense. To define 'mirroring process' in the narrow sense, we first need a definition of 'mirror neuron' or 'mirror system'. Rizzolatti, Fogassi, and Gallese (2004) offer the following definition of mirror neurons:

> Mirror neurons are a specific class of neurons that discharge both when the monkey performs an action and when it observes a similar action done by another monkey or the experimenter. (2004: 431)

This is a good definition of action mirror neurons but not mirror neurons in general. We don't want to restrict mirror neurons or mirroring processes to action-related events; they should equally be allowed in the domains of touch, pain, and emotion, for example. So let me propose a more general definition:

> Mirror neurons are a class of neurons that discharge both when an individual (monkey, human, etc.) undergoes a certain mental or cognitive event endogenously and when it observes a sign that another individual undergoes or is about to undergo the same type of mental or cognitive event.

One type of "sign" to which an observer's mirror neuron might respond is a behavioral manifestation of the mental event in question; another example is a facial expression. A third type of sign is a stimulus that can be expected to produce the mental event in question. Thus, an observer's pain mirror neurons discharge when he sees a sharp knife being applied to someone else's body.

The above definition of mirror neurons can be extended essentially unchanged to mirror systems or circuits:

> Mirror systems are neural systems that get activated both when an individual undergoes a certain mental or cognitive event endogenously and when he observes a sign that another individual is undergoing, or is about to undergo, the same type of mental or cognitive event.

What counts as an "endogenous" occurrence varies from one type of event to another. For present purposes I won't try to characterize endogenousness any further.

Even given the definitions of mirror neurons and mirror systems, it is not trivial to produce a definition of a mirroring process. One cannot say there is a mirroring process whenever a mirror neuron or mirror system discharges, or is activated. When a mirror neuron or system is endogenously activated, this is not a mirroring event. Only when a mirror neuron or system is activated in the observation mode is there a mirroring process. What exactly do we mean by "observation mode"? Must the observer perceive a genuine behavioral manifestation or expression (etc.) of an endogenous mirror event? I think not. If a good imitation of a pained expression triggers an observer's pain mirror-neuron to fire, the process in the observer is still a mirroring process although the imitation is not a genuine expression, or sign, of an endogenous mirror event.

It would also be incorrect to say, as a simple definition, that every non-endogenous activation of a mirror neuron or mirror system is a mirroring process. An important non-endogenous mode of activation is imagination-generated activation. Motor imagery, for instance, is the result of imagining the execution of a motor act, and the generation of motor imagery uses largely the same mirror circuits as used in the endogenous generation of an action (M. Jeannerod, personal communication). However, we would not consider the process of creating motor imagery a type of mirroring process. In light of these points, let us define a mirroring process as follows:

> Neural process N is an mirroring process if and only if (1) N is an activation of a mirror neuron or mirror system, and (2) N results from observing something that is normally a behavioral or expressive manifestation (or a predictive sign) of a matching mirror event in another individual.

3. Four Theses about Mirroring Processes and Mindreading

With these clarifications and definitions in hand, let me now present the main theses I wish to advance and defend in this chapter.

(T1) Mirroring processes in themselves do not constitute mindreading.

(T2) *Some* acts of mindreading ("low-level" mindreading) are caused by, or based on, mirroring processes.

(T3) Not *all* acts of mindreading (in particular, not "high-level" mindreading) are based on mirroring.

(T4) Simulation-based mindreading is broader than mirroring-based mindreading; some simulation-based mindreading (the "low-level" type) involves mirroring and some of it (the "high-level" type) doesn't.

In this section I'll defend thesis T1, and in succeeding sections theses T2, T3, and T4.

An act of mindreading consists of a belief or judgment about a mental state. So, if a mirroring process in itself were to *constitute* mindreading (as opposed to merely *cause* it), the "receiving" mirroring event would itself have to *be,* or *include*, a judgment or attribution of a mental state. In particular, it would have to be an attribution to a third person, presumably the originator of the mirroring process. Is there reason to suppose that belief "constitution" of this kind generally holds of mirroring processes?

Our definition of mirror neurons and mirror systems says that they are neural units that serve as substrates of one and the same cognitive event type, whether activated endogenously or observationally. This presumably implies that tokens of a mirror unit have substantially the same functional properties in whichever mode they are activated. If they had sharply different functional properties under different modes of activation, nobody would regard them as tokens of a mirroring type.[1] Having the same neuroanatomical location is not sufficient, because there are well-known cases in which the same neural region underpins more than one functionally distinct activity.[2] If that were the case in multi-modal cells or circuits, I doubt that anyone would speak of mirroring. So what are the cognitive or mental units that mirror neurons or mirror circuits underpin? They are units like "planning to grasp an object," "planning to tear an object," "feeling touch in bodily area X," "feeling pain in bodily area X," "feeling disgust", and so forth.

Now if the “receiving” mirror events are tokens of the same event types (i.e., they co-instantiate the same event types), then they too will be units like “planning to grasp an object,” “planning to tear an object,” “feeling touch in bodily area X,” and so forth. They won’t also be beliefs, judgments, or attributions to the effect that the observed agent is planning to grasp an object, planning to tear an object, feeling touch in bodily area X, and so forth. If they were beliefs, judgments, or attributions of these sorts (in addition to being plannings, feelings, etc.), then, since they are mirroring events, the original endogenous occurrences would also have to be beliefs, judgments, or attributions with the same contents. But nobody has ever proposed that the sending mirror events are, or include, beliefs, judgments, or attributions. These are strong considerations in favor of thesis T1.

The truth of T1 doesn’t spell doom for the idea that mirroring is pivotally involved in mindreading. T1 denies that a mirroring process in itself constitutes a mindreading event, but it allows a mirroring process to cause or generate a mindreading event. Lightning doesn’t constitute thunder, but lightning can certainly cause thunder. There is strong evidence for causal links between mirroring processes and selected mindreading events. This is thesis T2, which is addressed in the next several sections.

4. Mirroring and Intention Attribution

Since mirror systems were initially discovered in the domain of motor planning, one would reasonably expect this to be the domain for which mirror-based mindreading is best supported. Also, because the first proposal of a possible link between mirroring processes and mindreading (Gallese and Goldman, 1998) focused on the motoric domain, one might expect this domain to be favored as a locus of evidence for this connection. This is especially so in light of two recent studies concerning mirror-based intention attribution, one pertaining to monkeys and one to humans. I don’t agree entirely with the researchers’ own interpretations of their findings, but the second paper does provide plausible evidence to endorse their principal conclusion with respect to humans.

Fogassi et al. (2005) studied the discharge of monkey parietal mirror neurons during the viewing of a grasping act that would be followed by one of two subsequent acts: either bringing the object to the mouth or placing it in a container. Different grasping neurons of the viewer coded one or the other of the subsequent acts (or neither). In other words, for most of these grasping neurons the level of discharge was influenced by the subsequent motor act, although the discharge occurred in the observing monkey before the subsequent act began. Fogassi et al. write: “Thus, these [mirror] neurons not only code the observed motor act but also allow the observer to understand the agent’s intentions” (2005: 662). A similar moral is drawn from an experiment on humans by Iacoboni et al. (2005), viz., that the motor mirror system infers an agent’s intention. Now, to attribute an intention is to engage in mindreading. Do such intention attributions occur in virtue of the mirroring processes in and of themselves?

This is possible, but it isn’t definitely implied by the experimental evidence. There are two rival, comparatively “deflationary,” interpretations of the main findings that would not warrant this conclusion. The first rival interpretation would say that the parietal mirror neuron activity didn’t constitute the attribution of an *intention*, only the prediction of an *action*. The prediction of an action – since it’s not the attribution of a mental state – would not qualify as mindreading. The second rival interpretation would say that the parietal mirror neuron activity in the observer constituted a simulation, or mimicking, of the agent’s intention by the observer,[3] but not an intention attribution. *Possessing* an intention (or “tokening” an intention, as philosophers would say) should not be confused with attributing such an intention to the agent. Only such an attribution would be a belief or judgment *about* an intention.

Under either of these rival interpretations the reported scanning results do not, in isolation, definitively imply mindreading. The observing monkey or human that underwent mirroring processes during the experiment underwent *some* sort of mental events related to the conditions that discriminated between the different intentions of the agent. But on one interpretation the relevant mirror events in the observer were merely *action* predictions (hence not *mind*reading events); and on the other interpretation the mirror events were intention tokenings – again not mindreading events, this time because they were intentions rather than beliefs or attributions.[4]

If this were as far as the evidence goes, it wouldn’t firmly establish intention mindreading. But in fact there is additional evidence in the Iacoboni et al. study with human subjects, which had a special feature that lends support to intention attribution.

After being scanned, participants were debriefed about the grasping actions they had witnessed. They all reported that they associated the intention of drinking with the grasping action in the "during tea" condition and the intention of cleaning up with the grasping in the "after tea" condition. These verbal reports didn't depend on the instructions they had been given, i.e., whether or not they had been instructed to pay attention to the agent's intention. So here there is independent evidence of intention attribution (at least in humans). One highly probable scenario, then, is that human observers both mirrored the agent's intention by undergoing a matching intention themselves, and, in addition, these intentions were the causal bases for intention attributions to the agent. On this interpretation there is no suggestion that the observers' intentions constituted attributions; but it's agreed that intention attributions occurred.

Admittedly, the participants' reports during the debriefing session do not show that their intention ascriptions were caused by mirroring processes. But that is a reasonable inference, fully consistent with all findings. Even more open is the question of where their intention ascriptions occurred, whether in a motor mirroring area or elsewhere. Certainly the observers' (mirroring) intentions occurred in a motor mirroring area. But whether the attributions also occurred there is undetermined. However, an act of mindreading produced by a mirroring process need not have a mirroring process as its substrate. An attribution doesn't have to be part of a mirroring process; it only has to be caused by such a process. So the Iacoboni et al. study does provide support, if not conclusive support, for thesis T2.

5. Mirror-Based Attribution of Emotion

Support for thesis T2 finds additional fertile ground in the mindreading of emotion (commonly labeled emotion "recognition" in the studies to be reviewed). The best way to assemble the relevant evidence is to conjoin two sorts of studies: (1) studies of normal participants that establish the existence of mirroring processes for emotions, and (2) neuropsychological studies of emotion-specific brain damage showing that such damage is accompanied by selective impairment in attributing the same emotion. The two types of studies together yield convincing evidence that the substrate underpinning experience of an emotion is causally implicated in normal attribution of that emotion to an observed other. Failure to (fully) mirror an emotion in oneself while observing its expression in another prevents one from reliably attributing that emotion to the other.

The best example of this two-fold pattern of evidence pertains to disgust. Wicker et al. (2003) conducted an fMRI study of normal participants who were scanned both during the experience of disgust and during the observation of disgust-expressive faces. Participants viewed movies of individuals smelling the contents of a glass (disgusting, pleasant, or neutral) and spontaneously expressing the respective emotions. Then the same participants inhaled disgusting or pleasant odorants through a mask. The left anterior insula and the right anterior cingulate cortex were preferentially activated both during the experience evoked by inhaling disgusting odorants and during the observation of disgust-expressive faces. This establishes the existence of a mirroring process. However, the Wicker et al. study didn't feature any emotion recognition tasks, so the study didn't address the question of disgust attribution.

There are lesion studies, however, that contain relevant evidence about disgust attribution. Calder et al. (2000) studied patient NK who suffered insula and basal ganglia damage. In questionnaire responses NK showed himself to be selectively impaired in the experience of disgust (as contrasted with fear or anger). NK also showed significant and selective impairment in disgust recognition or attribution, which was established in two modalities: visual and auditory. Similarly, Adolphs et al. (2003) had a patient **B** who suffered extensive damage to the anterior insula (among other regions) and was able to recognize the six basic emotions except disgust when shown dynamic displays of facial expressions or told stories about actions. Apparently, an inability to mirror disgust because of damage to the anterior insula prevented these patients from attributing disgust, though their ability to attribute other basic emotions remained intact. It is a reasonable inference that when normal individuals recognize disgust when viewing the facial expression of disgust, this recognition is causally based on the production in the viewer of a (mirrored) experience of disgust.[5]

Analogous findings have been made in the case of fear, though here the results are a bit more qualified (Adolphs et al., 1994; Sprengelmeyer et al., 1999; Goldman and Sripada, 2005; Goldman, 2006: 115-116, 119-124; Keysers and Gazzola, 2006). Turning to the emotion of anger, Lawrence et al. (2002) reported selective anger recognition impairment as a result of “damage” to one of its “substrates.” Previous studies indicated that the neurotransmitter dopamine is involved in the experience of anger. Lawrence et al. therefore hypothesized that a temporary, drug-induced suppression of the dopamine system would also result in impairment of the recognition of angry faces while sparing recognition of other emotions. This is indeed what they found, though this has not been replicated in other studies.

Another finding in the emotion category concerns the secondary emotion guilt. According to the 1991 revised psychopathy checklist (PCL-R), psychopaths lack remorse or guilt. Blair et al. (1995) examined the ability of psychopaths and nonpsychopathic controls to attribute emotions to others, using a story understanding task. Responses of psychopaths and controls to happiness, sadness, and embarrassment stories did not significantly differ. But psychopaths were significantly less likely than controls to attribute guilt to others. This is indirect evidence, once again, that possessing the substrate of an emotion is critical to accurate attribution of that emotion to others, implicating a mirroring process as critical to normal attribution.

6. Pain and Touch

Continuing with thesis T2, is there evidence that mirroring plays a causal role in the (third-person) attribution of sensations like touch or pain? We start with touch. Keysers et al. (2004) showed that large extents of the secondary somatosensory cortex that respond to a subject’s own legs being touched also respond to the sight of someone else’s legs being touched. This is a clear demonstration of empathy for touch (at least in a minimal sense of ‘empathy’). However, there haven’t been tests to determine if observation-mediated somatosensory activity also causes attributions, or judgments, to the effect that another is undergoing such sensations.

Subsequent experiments do provide dramatic support for the mirroring-of-touch phenomenon, and even show that mirroring events can rise above the threshold of consciousness. Blakemore et al. (2005) described a subject **C** for whom the observation of another person being touched is experienced as tactile stimulation on the equivalent part of **C**’s own body. They call this vision-touch synaesthesia. fMRI experiments also reveal that, in **C**, the mirror system for touch (in both SI and SII) is hyperactive, above the threshold for conscious tactile perception. Banissy and Ward (2007) followed up on this study and confirmed that synaesthetic touch feels like real touch. However, neither of these studies specifically addressed the question of whether synaesthetic touch leads the subject to attribute the felt touch to the observed person, which would be interpersonal mindreading. Their findings are entirely consistent with this claim, but their experimental manipulations did not specifically address the question.

There is more evidence for mirroring-based pain attribution.   Mirror cells for pain were initially discovered serendipitously by Hutchison et al. (1999) while preparing a neurological patient for cingulotomy. More recently, Singer et al. (2004), Jackson et al. (2004), and Morrison et al. (2004) all reported pain resonance or mirroring. All three of these reports were restricted to the affective portion of the pain system, but subsequent transcranial magnetic stimulation (TMS) studies by Avenanti et al. (2005, 2006) highlighted the sensorimotor side of empathy for pain.

On the question of whether mirrored pain can cause pain attribution to others, results from both Jackson et al. (2004) and Avenanti et al. (2005) are especially pertinent. Jackson et al. had subjects watch depictions of hands and feet in painful or neutral conditions and were asked to rate the intensity of pain they thought the target was feeling. This intensity rating is a third-person attribution task. There was a strong correlation between the ratings (attributions) of pain intensity and the level of activity within the posterior ACC (a crucial component of the affective portion of the pain network). This confirms the idea that a mirror-induced feeling can serve as the causal basis of third-person pain attribution.

Avenanti et al. (2005; for a review, see Singer and Frith, 2005) found that there is sharing of pain between self and others not only in the affective portion of the pain system but also in the fine-grained somatomotor representations. When a participant experiences pain, motor evoked potentials (MEPs) elicited by TMS indicate a marked reduction of corticospinal excitability. Avenanti and colleagues found a similar reduction of corticospinal excitability when participants saw someone else receiving a painful stimulus, e.g., when participants watched a video showing a sharp needle being pushed into someone’s hand. No change in excitability occurred when they saw a Q-tip pressing the hand or a needle being pushed into a tomato. The neural effects were quite precise.

Corticospinal excitability measured from hand muscles was not affected by seeing a needle being thrust into someone's foot. Thus, there appears to be a pain resonance system that extracts basic sensory qualities of another person's painful experience and maps these onto the observers' own sensorimotor system in a somatotopically organized manner. Avenanti et al. also analyzed subjective judgments about the sensory and affective qualities of the pain ascribed to the model during needle penetration. These judgments were obtained by means of the McGill Pain Questionnaire (MPQ) and visual analogue scales, one for pain intensity and one for pain unpleasantness. The amplitude changes of MEPs recorded from the FDI muscle (the first dorsal interosseus) were negatively correlated with sensory aspects of the pain purportedly felt by the model during the 'Needle in FDI' condition, both for the Sensory scale of MPQ and for pain intensity on the visual analogue scale. Thus, judgments of sensory pain to the model seemed to be based on the mirroring process in the sensorimotor pain system. Finally, in a follow-up study, Avenanti et al. (2006) again found a significant reduction in amplitudes of MEPs correlated with the intensity of the pain being attributed to the model, and no MEPs modulation contingent upon different task instructions was found. In particular, specific sensorimotor neural responses did not depend on observers being explicitly asked to mentally simulate sensory qualities of others' sensations.

To sum up, there is adequate evidence in the case of pain to conclude that mirroring states or processes are often causally responsible for third-person mental attributions of the mirrored state, in further support of thesis T2.

7. The Limits of Mirror-Based Mindreading

There is clear evidence, then, that mirroring plays a causal role in certain types of mindreading. How wide a range of mindreading is open to a mirroring explanation? At present the range appears fairly narrow, because the types of mental activity known to participate in mirroring – viz., motoric activity, sensation, and emotion – appear to be circumscribed, although their ramifications for other phenomena are quite pervasive. Is it possible, then, that massive amounts of other mindreading are also based on mirroring? Or are there principled reasons to think that other types of mindreading differ from mirror-based mindreading?

Thesis T3 denies that there is a universal connection between mindreading and mirroring. A salient reason for being dubious of a universal connection comes from extensive fMRI studies of ‘theory of mind’ that identify different brain regions associated with desire and belief attribution, perhaps a dedicated mentalizing network. These regions are disjoint both from the well-known motoric mirror areas and from the areas involving pain and emotion that are cited above as loci of mirroring-based mindreading. The so-called ‘theory of mind’ regions are sometimes called “cortical midline structures”, and consist in the medial frontal cortex (MFC, perhaps subsuming the anterior cingulate cortex), the temporo-parietal junction, the superior temporal sulcus, and the temporal poles. Because these structures are strongly associated with mentalizing -- at least certain types of mentalizing -- there is a neuroanatomical challenge to the thesis that all mindreading is the product of mirroring. This is the “argument from neuroanatomy” for T3.

There are two challenges to this neuroanatomy argument for doubting the universality of mirroring in mindreading. The first challenge comes from the study of a stroke patient, **GT**, with extensive damage to the medial frontal lobes bilaterally, including regions identified as critical for ‘theory of mind’. Bird et al. (2004) carried out a thorough assessment of **GT**’s cognitive profile and found no significant impairment in ‘theory of mind’ tasks. They concluded that the extensive medial frontal regions destroyed by her stroke are not necessary for mindreading.

It is not clear, however, that medial prefrontal cortex should have been identified in the first place as the region dedicated to (high-level) mentalizing. Saxe and Wexler (2005) argue that the critical region is RTPJ (right temporo-parietal junction). If they are correct, then **GT** has no substantial bearing on the challenge to a universal role for mirroring in mindreading.

A more dramatic response to the neuroanatomy argument comes from the recent discovery of mirror neurons in the medial frontal lobe by Iacoboni’s group (Mukamel et al., 2007). Using single-cell recordings in humans, they found mirror cells for grasping and for facial emotional expressions in the medial frontal cortex, the sites being the preSMA/SMA proper complex, the dorsal sector of ACC, and the ventral sector of ACC. Iacoboni (personal communication) suggests that the findings may show that even the higher forms of mindreading are based on some mechanism of neural mirroring. Obviously, confirmation of the latter theory would undermine the argument from neuroanatomy. It remains to be seen, however, exactly which types of mindreading, if any, might be subserved by this new group of mirror cells.

Putting aside the argument from neuroanatomy, let us consider a second type of argument for the non-universality of mirroring-based mindreading: a theoretical argument that I'll call the "argument from error." This argument says that some forms of mindreading are susceptible to a form of error to which mirror-based mindreading isn't susceptible.[6] Therefore, not all mindreading is mirror-based. Let's spell this out.

Mirror-based mindreading is comparatively immune from error. In the first stage of mirror-based mindreading, the observer sees a behavioral or expressive sign in the agent that produces a matching mirror event in him. True, there might be a misfire here if the sign doesn't genuinely manifest a mental event of which it is typical. But this is not the kind of error I am thinking of in the case of "other" forms of mindreading. In the second stage of mirror-based mindreading, the mindreader classifies the mental event "received" from the agent and attributes it to the agent. If the classification process is normal, as well as the mirror-matching, the resulting act of mindreading will be accurate.

There are other types of mindreading, however, that are susceptible to a different kind of error. In particular, mindreading is prone to "egocentric" errors, largely from failures of perspective taking. Children are especially prone to this kind of error but it is also found in adults. One form of perspective-taking failure is the failure to inhibit self-perspective (for a review see Goldman, 2006: 164-175). There is no place for this kind of error in mirror-based mindreading. Thus, there must be a kind of mindreading that doesn't fit the mirroring mold.[7]

Notice that my point doesn't rest on the claim that mirror-based mindreading leaves no room at all for error. Recent evidence suggests that mirroring does not always guarantee matching, because it can be modulated by other information or preferences. Singer et al. (2006) found that empathic responses to pain are modulated by learned preferences. Participants played an economic game in which two confederates played fairly or unfairly, and participants then underwent functional imaging while observing the confederates receiving pain stimuli. Participants of both sexes exhibited mirrored pain responses, but in males the responses were significantly reduced when observing an unfair person receiving pain. If these mirror responses also generated pain attributions of varying levels, the indicated modulation would tend to produce errors. This is one way that mirror-based mindreading is open to error, but it's quite different from patterns of error found in other cases of mindreading.

A third argument for the non-universality of mirror-based mindreading is more straightforward than the first two. It is the simple point that a great deal of mindreading is initiated by imagination, and according to our definition of mirroring processes, imagination-driven events do not qualify as mirroring processes. Thus, if a person attempts to determine somebody else's mental state, not by observing their behavior or their facial or postural expression, but by learning about their situation from an informant's description, this act of mindreading will not involve a mirroring process. It may proceed by inference, by imagination, or by "putting oneself in the target's shoes," but none of these qualify as a mirroring process.

A fourth argument pertains to the types of mental states known to possess mirror properties. Most of these states are not states with propositional contents, like beliefs or desires; or if they do have propositional contents, these contents are of a bodily sort, pertaining to bodily location or bodily movement. Thus, states of pain and touch have mirror properties and the mirroring extends to their felt bodily locations. Intentions to act have mirror properties, and these intentions have contents concerning the types of effectors used (hand, foot, mouth) and the types of actions intended (coded in rich motoric terms). But there is no evidence that beliefs, for example, have mirror properties, especially beliefs with abstract contents. Observing another person grasp or manipulate an object with his hands elicits in the observer a covert intention to grasp or manipulate an object. But observing someone else who is reflecting on the problem of global warming does not elicit a similar thought in one's own mind (except by sheer coincidence). Beliefs and other reflective states do not elicit matching contentful states by a mirroring process; nor do desires that go unexpressed in a distinctive motoric signature. Thus, there is a large class of mental states that aren't mirrored. Since they surely are the targets of mindreading, they must be read in a different fashion. This establishes thesis T3.

8. High-Level Simulation-Based Mindreading

I turn finally to thesis T4. Many writers equate mirroring with simulation. So if a given mental state cannot be read by a mirroring process, it cannot be read by simulation. I take a different view (Goldman, 2006). Simulation and mirroring are not equivalent; mirroring is just one species of simulation. Hence, if a type of mental state isn't readable by mirroring, it's still possible it can be read by simulating, just a different form of simulating. It's also possible, of course, that it can be read by theorizing, and I don't wish to deny that some acts of mindreading, partly or wholly, consisting of theorizing (Goldman, 2006: 43-46). Here I shall focus on the second form of simulational mindreading.

The basic idea of simulation of this second kind is to "re-enact" or "re-create" a scenario in one's mind that differs from what one currently experiences in an endogenous fashion. It is to imagine a scenario, not merely in the sense of "supposing" that it has occurred or will occur, but to imagine being immersed in, or witnessing, the scenario. In other words, it involves engaging in mental "pretense" in which one tries to construct the scenario as one would experience or undergo it if it were currently happening. This is what philosopher-simulationists had in mind originally by "simulation" (Gordon, 1986; Heal, 1986; Goldman, 1989, 2006; Currie and Ravenscroft, 2002), not mirroring, which is a more recent entrant onto the scene (Gallese and Goldman, 1998). Mirroring features an automatic re-creation in an observer's mind of an episode that initially occurs in another's mind. In "enactment simulation," by contrast, one attempts to create such a matching event without currently observing another person who undergoes it. One tries to construct the event with the help of experience or knowledge that, it is hoped, will facilitate the construction. Successful re-enactment or re-creation is more problematic than accurate mirroring. Re-enactment must typically be guided by knowledge stored in memory, the quality of which is quite variable. However, any attempt at re-enactment can be called "simulation" whether or not there is successful, accurate matching (Goldman, 2006: 38).

Enactment simulation as sketched here approximates the notion of simulation evoked by the neuroscientists Buckner and Carroll (2007), who discuss it under the heading of “self-projection.” They conceive self-projection as the mental exploration and construction of alternative perspectives to one’s current actual perspective, including perspectives on one’s own future (“prospection”), one’s own past (autobiographical memory), the viewpoint of others (theory of mind), and navigation. Buckner and Carroll refer to imagining an alternative perspective as “simulation.” They also argue that all these forms of self-projection involve a shared neural network involving frontal and medial temporo-parietal lobe systems that are traditionally linked to planning and episodic memory.

Buckner and Carroll cite a variety of evidence in support of their view, beginning with the fact that among the deficits created by frontal lobe lesions are deficits in planning and structuring events in an appropriate temporal sequence. Patients with frontal lesions often perform normally in well-established routines and can show high intellectual function, but when confronted with challenging situations and new environments, reveal an inability to plan. They are unable to order sequences temporally, plan actions on tasks requiring foresight, and adjust behaviors flexibly as rules change. Mesulam (2002) noted that the prefrontal cortex might have a pivotal role in the ability to “transpose the effective reference point [of perception] from the self to other, from here to there, and from now to then.”

Other evidence concerns the medial temporal lobe, damage to which often causes amnesia. A lesser-studied aspect of the amnesic syndrome is the inability to conceive the personal future. In his seminal description of amnesia in Korsakoff’s syndrome, Talland (1965) noted that his amnesic patients could say little about their future plans. The same was true of the amnesic patient HM. Similarly, Klein et al. (2002) observed that their amnesic patient DB, when questioned about his future, either confabulated or did not know what he would be doing. Although DB had general knowledge of the future -- he knew there was a threat of weather changes -- he lacked the capacity to consider himself in the future.

A propos of theory of mind, Buckner and Carroll draw on Gallagher and Frith’s (2003) account of the role of frontopolar cortex. They suggest that the paracingulate cortex, the anterior-most portion of the frontal midline, is recruited in executive components of simulating others’ perspectives. This region is contiguous with but distinct from those involved in episodic remembering. Gallagher and Frith also conclude that this region helps to "determine [another's] mental state, such as a belief, that is decoupled from reality, and to handle simultaneously these two perspectives on the world." Obviously, this is the kind of ability crucial in solving false-belief tasks in mindreading.

If Buckner and Carroll are right that a (substantial sector) of mentalizing activities are simulations that conform to the foregoing description, and if they are right that such activities take place (roughly) in the brain systems they identify, then it appears that these are not mirroring activities. Nonetheless, they are simulation activities, in the sense intended Thus, a substantial chunk of mindreading is simulationist in character without being the product of mirroring. In Simulating Minds (Goldman, 2006) I distinguish two types of simulation for mindreading: "low-level" and "high-level". Low-level simulation features mirroring and high-level simulation does not. Simulating Minds does not try to pinpoint precisely all the brain regions associated with high-level mindreading, and that is not essential here either. What is interesting about Buckner and Carroll's contribution is that it identifies a certain network or circuit of brain regions that accomplish a certain general type of task (adopting an alternative perspective), which is instantiated in other domains as well as mindreading. This tends to substantiate thesis T4.

9. Interactions between Cortical Midline Structures and Mirror Systems?

Uddin et al. (2007) propose a unifying model to account for data on self and social cognition by sketching links between cortical midline structures (CMS) and the (motor) mirror-neuron system (MNS). The former is taken to consist of the medial prefrontal cortex, the anterior cingulate cortex and the precuneus, and the latter is composed of the inferior frontal cortex and the rostral part of the inferior parietal lobule. They argue that a right-lateralized frontoparietal network that overlaps with mirror-neuron areas seems to be involved with self-recognition and social understanding. Because both MNS and CMS are involved in self-other representations, it seems only natural, they propose, that the two systems interact. One pathway by which this might occur is a direct connection between the precuneus (which they regard as a major node of the CMS) and the inferior parietal lobule (the posterior component of the MNS). Also there are direct connections between mesial frontal areas and the inferior frontal gyrus. Thus, the anterior and posterior nodes of the CMS and MNS are in direct communication.

However, it seems that MNS and CMS perform quite different functions vis-à-vis self-understanding. Both the self-face and the self-body activate the right frontoparietal network (Uddin et al. 2005; Sigiura et al. 2006). So the right-lateralized system, associated with the mirror-neuron system, seems to be related to representations of the physical self rather than the mental self. CMS structures, on the other hand, seem to be more involved in internal aspects of representing self and others, including mentalizing, as Uddin et al. (2007) themselves concede.

Uddin et al. (2007) propose a division of labor in which the CMS might support "evaluative simulation" in the same way that the MNS supports "motor simulation." This division of labor between the two networks would yield specializations for two related processes that are crucial to navigating the social world: understanding physical actions of intentional agents and understanding the attitudes of others. It is unclear, however, exactly what they mean by "evaluative simulation." Not all the mentalizing work done by the CMS involves evaluation in any straightforward sense. Attributing beliefs to other people (including false beliefs) is a principal mentalizing activity executed by portions of the CMS. But there is nothing "evaluative" (as opposed to "descriptive") about a belief attribution; nor are beliefs themselves evaluative states. It is also unclear what Uddin et al. (2007) mean by "simulation" in this context; they offer no explanation of this (somewhat slippery) notion. However, it appears that we agree on two important points: that simulation plays a central role in different sectors of mentalizing and that mirror-neuron systems perform only a portion, albeit a very fundamental portion, of the mentalizing work that the human mind undertakes.

10. Conclusion

Mirroring per se does not constitute mindreading. Nonetheless, there is evidence of mirroring-based mindreading in several domains, including action intention, emotion, and pain. Mirroring-based mindreading is what I call "low-level" mindreading. There are also many reasons, however, to doubt that all mindreading is based on mirroring. How does this bear on the simulation theory of mindreading? Mirroring is one kind of simulational process but not the only one. Attempting to take another person's perspective, or put oneself in their shoes, is another type of simulational process, and this kind of process is extensively used in mindreading. Thus, simulation figures importantly in "high-level" as well as "low-level" mindreading.

Acknowledgements   The author thanks Vittorio Gallese, Marco Iacoboni, and Frederique de Vignemont for detailed comments that resulted in many helpful changes in the manuscript. Other helpful comments were due to Holly M. Smith.

References

Adolphs, R., Tranel, D. and Damasio, A. R. (2003). Dissociable neural systems for recognizing emotions. Brain and Cognition 52: 61-69.

Adolphs, R., Tranel, D., Damasio, H. and Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the amygdale. Nature 372: 669-672.

Avenanti, A., Bueti, D., Galati, G. and Aglioti, S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. Nature Neuroscience 8: 955-960.

Avenanti, A., Paluello, I. M., Bufalari, I. and Aglioti, S. M. (2006). Stimulus-driven modulation of motor-evoked potentials during observation of others' pain. NeuroImage 32: 316-324.

Bartels, A. and Zeki, S. (2000). The architecture of the colour centre in the human visual brain: New results and a review. European Journal of Neuroscience 12: 172-193.

Bird, C. M., Castelli, F., Malik, O., Frith, U. and Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. Brain 127: 914-928.

Banissy, M. J. and Ward, J. (2007). Mirror-touch synaesthesia is linked with empathy. Nature Neuroscience 10: 815-816.

Blair, R. J. R., Sellars, C., Strickland, I., Clark, F., Williams, A. O., Smith, M. and Jones, L. (1995). Emotion attributions in the psychopath. Personality and Individual Differences 19: 431-437.

Blakemore, S.-J., Bristow, D., Bird, G., Frith, C. and Ward, J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. Brain 128: 1571-1583.

Buckner, R. L. and Carroll, D. C. (2007). Self-projection and the brain. Trends in Cognitive Sciences 11(2): 49-57.

Calder, A.J., Keane, J., Manes, F., Antoun, N., and Young, A.W. (2000). Impaired recognition and experience of disgust following brain injury. Nature Reviews Neuroscience 3: 1077-1078.

Currie, G. and Ravenscroft, I. (2002). Recreative Minds, Imagination in Philosophy and Psychology. Oxford: Oxford University Press.

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

Decety, J. and Chaminade, T. (2005). The neurophysiology of imitation and intersubjectivity. In S. Hurley and N. Chater, eds., Perspectives on Imitation: From Neuroscience to Social Science (pp. 119-140). Cambridge, MA: MIT Press.

De Vignemont, F. and Haggard, P. (in press). Action observation and execution: What is shared? Social Neuroscience.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F. and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. Science 308: 662-667.

Gallagher, H. L. and Frith, C. D. (2003). Functional imaging of 'theory of mind'. Trends in Cognitive Sciences 7: 77-83.

Gallese, V. (2005). "Being-like-me": Self-other identity, mirror neurons, and empathy. In S. Hurley and N. Chater, ed., Perspectives on Imitation, vol. 1 (pp. 101-118). Cambridge, MA: MIT Press.

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

Gallese, V. and Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 2: 493-501.

Gallese, V., Keysers, C. and Rizzolatti, G. (2004). A unifying view of the basis of social cognition. Trends in Cognitive Sciences 8: 396-403.

Goldman, A. I. (1989). Interpretation psychologized. Mind and Language 4: 161-185.

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

Goldman, A. I. and Sripada, C. R. (2005). Simulationist models of face-based emotion recognition. Cognition 94: 193-213.

Gordon, R. M. (1986). Folk psychology as simulation. Mind and Language 1: 158-171.

Heal, J. (1986). Replication and functionalism. In J. Butterfield, ed., Language, Mind and Logic. Cambridge: Cambridge University Press.

Hutchison, W. D., Davis, K. D., Lozano, A. M., Tasker, R. R. and Dostrovsky, J. O. (1999). Pain-related neurons in the human cingulate cortex. Nature Neuroscience 2(5): 403-405.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziota., J.C. et al. (1999). Cortical mechanisms of human imitation. Science 286: 2526-2528.

Iacoboni, M. (2005). Understanding others: imitation, language, and empathy. In S. Hurley and N. Chater, eds., Perspectives on Imitation: From Neuroscience to Social Science (pp. 76-100). Cambridge, MA: MIT Press.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C. and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. PLoS Biology 3: 529-535.

Jackson, P. L., Meltzoff, A. N. and Decety, J. (2004). How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage 24: 771-779.

Keysers, C. and Gazzola, V. (2006). Towards a unifying neural theory of social cognition. Progress in Brain Research 156: 383-406.

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

Klein, S. B. et al. (2002). Memory and temporal experience: the effect of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. Social Cognition 20: 353-379.

Lawrence, A.D., Calder, A.J., McGowan, S.M. and Grasby, P.M. (2002). Selective disruption of the recognition of facial expressions of anger. NeuroReport 13(6), 881-884.

Mesulum, M. M. (2002). The human frontal lobes: transcending the default mode through contingent encoding. In D. T. Stuss and R. T. Knight, eds., Principles of Frontal Lobe Function (pp. 8-30). Oxford: Oxford University Press.

Morrison, I., Lloyd, D., de Pelligrino, G. and Robets, N. (2004). Vicarious responses to pain in anterior cingulate cortex. Is empathy a multisensory issue? Cognitive Affective Behavioral Neuroscience 4: 270-278.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M. and Fried, I. (2007). Mirror properties of single cells in human medial frontal cortex. Social Neuroscience Abstracts.

Rizzolatti, G. (2005). The mirror neuron system and imitation. In S. Hurley and N. Chater, ed., Perspectives on Imitation, vol. 1 (pp. 55-76). Cambridge, MA: MIT Press.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience 27: 169-192.

Rizzolatti, G., Fogassi, L. and Gallese, V. (2004). Cortical mechanisms subserving object grasping, action understanding, and imitation. In M. Gazzaniga, ed., The Cognitive Neurosciences III (pp. 427-440). Cambridge, MA: MIT Press.

Saxe, R. (2005). Against simulation: The argument from error. Trends in Cognitive Sciences 9: 174-179.

Saxe, R. and Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. Neuropsychologia 43: 1391-1399.

Sigiura, M. et al. (2006). Multiple brain networks for visual self-recognition with different sensitivity for motion and body part. Neuroimage 32: 1905-1917

Singer, T. and Frith, C. (2005). The painful side of empathy. Nature Neuroscience 8: 845-846.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., and Frith, C. (2004). Empathy for pain involves the affective but not sensory components of pain. Science 303: 1157-1162.

Singer, T., Seymour, B., O'Doherty, J. Stephan, K. E., Dolan, R. J. and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature 439: 466-469.

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T. and Przuntek, H. (1999). Knowing no fear. Proceedings of the Royal Society, series B: Biology 266: 2451-2456.

Talland, G. A. (1965). Deranged Memory: A Psychonomic Study of the Amnesic Syndrome. New York: Academic Press.

Uddin, L. Q. et al. (2005). Self-face recognition activates a frontoparietal 'mirror' network in the right hemisphere: an event-related fMRI study. Neuroimage 25: 926-935.

Uddin, L. Q., Iacoboni, M., Lange, C. and Keenan, J. P. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. Trends in Cognitive Sciences 11(4): 153-157.

Wicker, B., Keysers, C., Plailly, J., Royet, J-P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. Neuron 40: 655-664.

[1] The functional properties of two mirror tokenings need not be identical, however. First, it is taken for granted that mirror discharges in execution and observation mode are not perfectly identical (for a review, see Csibra, 2007). In observation mode the frequency or amplitude of firing may not coincide with that of the execution mode. Thus, the "strength" of two tokenings may diverge slightly, with implications of slight differences in functional properties. Second, the Parma group from the beginning has distinguished between "strictly" and "broadly" congruent mirror neurons (Gallese et al. 1996). In the case of broad congruence, functional properties are not identical. For present purposes, however, we can ignore this issue. Our approach focuses, for simplicity, on strictly congruent mirror neurons (or their analogue in mirror systems or circuits).

[2] For example, lesions to the fusiform gyrus of the right occipital lobe produce both prosopagnosia and achromatopsia (Bartels and Zeki 2000). But these two deficits have no interesting functional relationship to one another. It just so happens that the impaired capacities are at least partially co-localized in the fusiform gyrus.

[3] I usually speak of the simulation relation as holding between processes rather than states (including intentions). However, as I use the term, a process is a series of causally related states; so, as a limiting case, we may consider a state to be a process with a single member. Hence, we may also speak of states, such as intentions, as items that figure in simulation relations.

[4] De Vignemont and Haggard (in press) make a strong case for the claim that the best candidate for what is shared in a pair of mirroring events is an "intention in action." If this is right, it argues against the intention-prediction interpretation of the Iacoboni et al. (2005) imaging results per se.

[5] It is assumed in all of these studies that the participant not only "recognizes" the emotion in the sense of classifying or categorizing it, but also views the emotion as occurring in the observed target (whose facial expression is shown or depicted). This implies that the participant is not merely categorizing the emotion but also attributing it to the target. If the categorization results from the mirroring process -- which includes the observation of the target -- it is hardly open to question that the attribution also results from the mirroring process. Thanks to F. de Vignemont for emphasizing this point.

[6] Saxe (2005) uses a somewhat analogous argument from error to criticize the general simulation theory of mindreading. Here an argument from error is being used to resist the claim that all mindreading takes a specific simulationist form, viz. mirroring-based mindreading. Many errors associated with non-mirror-based mindreading are readily accommodated by a second form of simulation, discussed below in section 8. More generally, see Goldman (2006: chap. 7).

[7] It might be replied that mirror-based mindreading is susceptible to egocentric error. F. de Vignemont (personal communication) suggests that if I myself have a terrible back pain and I see you carrying a heavy box, I would feel pain and ascribe this feeling to you. This might be an error because you are perfectly fine with a box that heavy; you are not in pain. Isn't this an egocentric error? No doubt, it is an egocentric error. The question is whether it's a case of mirroring, at least a pure case of mirroring. It isn't a case in which I see you exhibiting a behavioral or expressive manifestation of pain. And it's questionable whether the perceived heaviness of the box is a "sign" of pain comparable to a knife or needle penetrating a body. It might be a case of inference- or imagination-caused pain rather than mirror-produced pain Admittedly, the case puts pressure on our definition of mirroring, but this isn't a problem only for me. It's a problem for anyone seeking to be precise about what counts as mirroring. In any case, there are other arguments offered here in favor of thesis T3. It doesn't rest exclusively on the lesser-liability-to-error argument.