

Jacob on Mirroring, Simulating and Mindreading

Alvin I. Goldman

Rutgers University

Abstract

Jacob (2008) raises several problems for the alleged link between mirroring and mindreading. This response argues that the best mirroring-mindreading thesis would claim that mirror processes cause, rather than constitute, selected acts of mindreading. Problems with a constitution thesis do not carry over to a causation thesis. Second, the best current evidence for mirror-based mindreading is found not in the motoric domain but in the domains of emotion and sensation, where the evidence (ignored by Jacob) is substantial. Finally, simulation theory should distinguish low-level simulation (mirroring) and high-level simulation (featuring pretence or imagination). Jacob implies that bi-level simulationism creates an unbridgeable “gap” in intention reading, but this is not a compelling challenge.

I

Pierre Jacob (2008) raises a number of interesting challenges to the thesis that mirror processes play an important role in mindreading. Proponents of this thesis need to clarify the conceptual and empirical bases for their claims. As Jacob indicates, different proponents of the mirroring-mindreading thesis (henceforth, MM thesis) approach it in somewhat different ways and interpret the existing evidence slightly differently. The present response to his challenges reflects my own theoretical perspective only, and doesn't correspond precisely to the perspective of the Parma group (the discoverers of

mirror phenomena) and their collaborators. But there is room for multiple theoretical and interpretive perspectives. From my perspective it is important to consider domains of mirroring and mindreading that Jacob sets aside, the domains of emotion and sensation. Evidence from these domains is more telling than evidence from motor mirroring and motor intentions.

Jacob and I agree on the fundamental assumption that mindreading consists of attributing (ascribing, imputing) a mental state to someone, for present purposes, another person. The state attributed must be mental rather than merely behavioral; otherwise the attribution doesn't qualify as mindreading. Furthermore, to attribute a mental state to an individual is to represent that individual as being in that state. There may be debate about whether the representation must be "conceptual" rather than "nonconceptual," but some sort of representation of the target as being in a mental state is essential for mindreading. Given this shared starting point, there is the question of how mirroring relates to mindreading.

What is mirroring? I shall offer two definitions: the first aims to capture the "core" meaning of mirroring and the second to express the more general and relaxed sense.

(M_C) A neural process or event E is an instance of core-mirroring just in case E is the activation in an observer of a neuron or neural system that (1) results from observing a target's behavior or behavioral expression and (2) matches or replicates an activation in the target of a corresponding neuron or neural system that the observed behavior or expression manifests.

(M) A neural process or event E is a case of mirroring just in case E is the activation in an observer of a neuron or neural system that (1) results from observing a target's behavior or behavioral expression and (2) would, in a normal case of such behavior, match or replicate an activation in the target of a corresponding neuron or neural system that the observed behavior would manifest.

Core-mirroring captures the fundamental idea of mirroring, the idea of interpersonal neural matching, replication, or duplication (within some selected parameters). I suspect, however, that the common use of 'mirroring' is not constrained by the original guiding idea. Consider a film featuring a virtual human. It looks like an ordinary person, but is in fact the creation of animators. If an observer of the film undergoes the same sorts of neural mirroring responses to the character's actions and emotional expressions that she would undergo if the character were real, surely these would be considered mirror processes, despite the fact that no brain processes in the observer literally match or replicate any corresponding brain processes in the target -- because the target has no brain. This indicates that core-mirroring, as defined above, is not the concept used in scientific practice. The concept in use is best captured by (M), which is neutral about the existence of a target whose brain activities are actually matched or replicated. Of course, core-mirroring is the basic and original concept of mirroring, but it's more demanding than necessary. For most of the following discussion, the distinction between mirroring and core-mirroring won't be important. In only one case will it be of interest.

How does mirroring relate to mindreading? A mirroring activation might relate to mindreading in one of two ways. First, it might constitute an instance of mindreading.

That is, a mirroring activation might be an attribution to a target of a specific mental state. Second, the neural activation might cause (or causally contribute to) a distinct neural event or set of events that constitute an attribution of a mental state to the target.¹ Jacob sometimes writes as if the first interpretation is the only, or the preferred, interpretation of the MM thesis. This is definitely not the only interpretation of the MM thesis, and it's not the one I prefer. The second interpretation of the MM thesis offers a more promising position, as argued in (Goldman, 2006) and below. Under the second interpretation it would suffice for proponents of MM to show that mirror events generate other events or states that constitute mental-state representations. Similar remarks apply to the so-called "simulation" approach to mindreading, which bears a close relation to MM. (This relation requires careful delineation; see section IV.) By my lights, the best way to spell out a simulation approach is to say that mindreading events, or a substantial subset of them, are caused by episodes of mental simulation.

Is my (currently) preferred interpretation of MM identical to the form of MM suggested by Gallese and Goldman (1998)? Jacob devotes considerable attention to this paper, presumably because it was the first to propose a mirroring-mindreading link.

Gallese and I were not entirely explicit on the constitution/causation issue. We wrote:

Here we explore [the] possibility ... that MNs underlie the process of 'mind-reading', or serve as precursors to such a process.... MNs are part of ... the folk psychologizing mechanism" (1998: 495).

Because being part of a mindreading process or mechanism allows for the possibility that other parts perform the mental-state attribution, this language made no commitment to the constitution interpretation; but neither was it committed to the causal interpretation.

According to the simulation theory (ST) of mindreading as we presented it, mindreading is

... an attempt to replicate, mimic, or impersonate the mental life of the target ...

[It is a] process of mimicking (or trying to mimic) the mental activity of the target agent. (1998: 497).

The mimicking part of a mindreading process, however, might precede the part that performs the attribution. It might cause rather than constitute the attribution. Thus, a mirror theorist could favor ST without endorsing the idea that mirroring itself qualifies as mental-state attribution.

In some passages of Jacob's paper, he seems to ascribe the constitution interpretation to the Gallese and Goldman paper, or comes close to doing so. He then uses this interpretation as the basis for a variety of criticisms. Here is the first passage where Jacob seems to impute the constitution interpretation to us:

... [B]y performing a mental simulation of the agent's observed movements, the activity of MNs is seen as enabling the observer to recognize the agent's action, or even to represent her intention (or goal). Since representing an agent's intention is unquestionably part of third-person mindreading, it turns out that one fundamental function of MNs is to underlie mindreading. (Jacob, 2008: pp. 3-4 of PDF)

The first sentence of the passage speaks of MNs as "enabling" the observer to represent the agent's intention (or goal), a verb that might be compatible with a causal interpretation but is more plausibly read here in constitution terms. The constitution interpretation is clearer in the second sentence, which speaks of MNs as having the

function of “underlying” mindreading, which suggests that MNs are the neural substrate of mindreading. Two paragraphs later Jacob lodges the first of his objections to the MM thesis. “[I]t is highly questionable whether by mentally rehearsing an agent’s observed movements, an observer could represent the agent’s underlying intention” (Jacob, 2008: p. 4 of PDF).² In challenging the view that mental rehearsal, i.e. mirroring, could represent the agent’s intention, the passage presupposes the constitution interpretation. As indicated, however, this is not a construal of MM that proponents need to accept. In particular, I would not accept it.

On the other hand, members of the Parma team probably lean toward this construal, so Jacob is well within his rights to discuss problems for this type of view. He quotes a passage from Gallese et al. (2004) to the effect that mirror mechanisms allow an observer to have a “direct experiential grasp of the mind of others”. Gallese et al. further say that mirroring directly links ‘I do and I feel’ with ‘he does and he feels’. Grasping the mind presumably means grasping a mental state, and linking ‘I feel’ with ‘he feels’ seems to imply that one attributes ‘he feels’ to the target individual. Here it looks indeed as if mental-state attribution is considered an activity performed by mirroring mechanisms.

I would agree with Jacob in questioning this view of mirroring activity. Restricting attention to motor mirroring for the moment, what type of mental state or event is instantiated by the activity of primary interest in the premotor region of an agent’s brain? Presumably, it is motor plans or intentions, i.e. propositional attitudes with contents of the form “let [my] effector E perform motor act M with respect to goal-object G.” We can notate the content of such an event as follows: <effector E, motor act

M, goal-object G>. If something like this construal is right, then premotor activation in the execution mode has representational content. But does its representational content include mentalistic content? Apparently not. A premotor planning state is a mental state (a plan or intention) that has a certain content, but this content does not include any mental-state constituents. Intends, for example, is not part of its content. Analogously, if someone believes “2+2 = 4,” the content of the belief is purely arithmetic and wholly non-mentalistic, although the belief that bears the arithmetic content is a mental state. To take another example, if a person verbally asserts the proposition that 2+2 = 4, the act of assertion is a speech act, but the content of the speech act contains nothing concerning speech.

Now, if an observer’s mental state mirrors that of an agent, and is therefore congruent with it, the observer presumably instantiates a state with the same content, <effector E, motor act M, goal-object G>. Once again, however, intending will not be part of its content. By contrast, a mental act of attributing an intention to another person (an act of mindreading) would be a belief that features the concept intends in its content. The entire content of such a belief might be notated in roughly the following way: <person A, intention I, time t>. This content is not congruent with the presumptive content of the motor plan in the agent’s brain. Moreover, the state is a belief rather than an intention. So it looks implausible that motor mirroring events are states of attribution (beliefs) containing mentalistic contents. This explains some of my own reasons for doubting the correctness, or fruitfulness, of what I take to be the Parma approach to the issue.³

In appraising these arguments, it is important not to conflate “goal” and “intention.” The Parma team often stresses that motor MNs code for the goals of observed actions. This may be uncontroversial if “goal” means “goal object,” which might refer either to a physical object (e.g. a cup) or a physical event or outcome (e.g. a cup being moved to one’s mouth). But the same term “goal” can also be used to refer to a mental state, a state of desire with a certain intentional object or relation of “aboutness.” In this sense it is not uncontroversial that MNs code for goals -- especially monkey MNs. Thus, the following passage from Fogassi et al. (2005) contains a very debatable transition:

[T]his neuronal property allows the monkey to predict what is the goal of the observed action and, thus, to ‘read’ the intention of the acting individual.

(Emphasis for ‘thus’ is mine)

The term ‘thus’ involves the questionable transition. From the fact that a goal-event -- a non-mental event -- is predicted it does not follow that a mental event such as an intention is predicted. In defending a constitution interpretation of the MM thesis, it is important to keep these different senses of the term “goal” apart from one another.

To repeat, the foregoing arguments do not have the implication that mirroring plays no role in mental attribution. They are entirely compatible with the thesis that mirroring (sometimes, often, commonly) causes a person to attribute a motor intention to an observed agent. Moreover, if mirroring causes mindreading in a simple case where one attributes a motor intention (an “intention in action”) based on an action visibly executed just now, then it’s an instance of retrodictive attribution.

Is the causal form of the MM thesis correct, however? Does mirroring cause mindreading? What evidence supports this thesis? The Gallese and Goldman paper was a speculative paper; it sought to construct a theoretically plausible link between mirroring and mindreading. It did not adduce specific experimental evidence beyond the early mirror-neuron findings, which themselves did not explicitly address mindreading. Mirror-neuron findings were soon extended, however, to experiments on additional topics in social cognition, for example, imitation (Iacoboni et al., 1999). The first substantial studies addressed to mirroring and motor mindreading were (Fogassi et al. 2005) and (Iacoboni et al., 2005), two closely parallel studies, although the first reported an experiment with monkeys and the second an experiment with humans. The two papers had a slightly different focus than the “standard” case of retrodictive attribution.⁴ They dealt with intentions to perform future acts, beyond those the attributer currently observes. Moreover, as Iacoboni et al. explain, an analysis of the role of mirror neurons in these scenarios requires the postulation of a different class of mirror neurons, distinct from classical MNs. Jacob devotes part of his critique to this evidence and its theoretical underpinning, so let us examine these findings and their purported explanations in closer detail.

II

Iacoboni et al. (2005) performed a fMRI experiment in which subjects observed video clips presenting three kinds of stimulus conditions: grasping hand actions without any context (Action condition), scenes specifying a context without actions (a table set for drinking tea versus ready to be cleaned up after tea) (Context condition), and grasping

hand actions performed in either the before-tea or after-tea context (Intention condition). In the Intention condition, the context suggested a further intention beyond that of grasping a cup: an intention either to drink tea or to clean up (after tea). The Intention condition yielded a significant signal increase in premotor mirroring areas where hand actions are represented. The investigators interpreted this as evidence that premotor mirroring areas are involved in understanding the intentions of others, not only motor intentions for acts currently being observed but intentions for subsequent actions. In other words, the observer's mirror neurons code for an actor's intention to do a sequence or chain of actions in which the currently observed motor act is embedded.

Jacob raises two problems for the proffered explanation of these findings. His first worry is whether these experiments genuinely show that the MN activity generates – that is, causes -- a representation of the agent's "prior" intention, i.e. an intention to do a subsequent part of a larger action. The worry is that both of the experiments reveal correlations between enhanced MN activity and the facilitation of a representation of an agent's prior intention, but correlation is not causation. Instead of being caused by MN activity, an observer's representation of the agent's prior intention might be caused, for example, by perceptual processing of contextual cues.

This is an issue I'll let others address (for example, the original investigators). Perhaps additional experimental controls are needed to exclude the alternative explanations Jacob suggests. Or perhaps background knowledge of certain properties of these networks might make his alternative hypotheses implausible. I won't try to resolve this issue.

However, I have worries of my own about the Iacoboni et al. interpretation of their findings (see Goldman, 2008a). There are two rival, comparatively “deflationary” interpretations of the findings that would not warrant the conclusion that MN activity is responsible for intention attribution. The first rival interpretation would say that the enhanced activity in MN areas (during observation of the Intention condition) did not constitute the attribution (prediction) of an intention, but only the prediction of an action. Since an action is not a mental state, predicting an action would not qualify as mindreading. The second rival interpretation would say that what occurs in the relevant mirror area of the observer is a mimicking of the agent’s intention but not an intention attribution (belief). Possessing or tokening an intention 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.

Jacob’s second worry is whether the experimentally obtained responses are really the product of mirroring or resonance, even if they are the result of precisely the kind of neural activity that the investigators themselves invoke. This problem centers on a distinctive feature of the present studies. Iacoboni et al. contend that mirror neurons are the likely neurons driving the signal changes they found. They concede, however, that these neurons could not be the standard, or “classical”, type of mirror neurons, that is, neurons that normally display congruence between their visual and motor properties. The congruence property cannot account for the differences in observed response between the drinking and cleaning Intention clips. The same grasping action is observed in both cases, so any standard mirroring response ought to be the same, not different. However, another set of neurons in inferior frontal cortex were previously reported and referred to

as “logically related” neurons (di Pellegrino et al., 1992). These neurons are visually triggered by a given motor act, e.g., the observation of grasping, but discharge during the execution not of the same motor act (grasping), but of another act, functionally related to the observed one (e.g. bringing to the mouth).

Jacob argues that the new model based on chains of logically related (really, probabilistically related) MNs abandons the assumption of strong congruence between the motor and perceptual properties of MNs. MNs in an observer’s brain do not, strictly speaking, resonate with the concurrent MNs in an agent’s brain. While the agent’s MNs control the execution of an act of grasping, the observer’s MNs control an act of drinking. Acceptance of the new model based on chains of “logically related” MNs generates a discrepancy between motor contagion and the activity of MNs. Jacob is asking, in effect: How does this story about logically related neurons support the thesis that intention attribution is executed by mirroring? Logically related neurons don’t really mirror.

This is an important point. It challenges the claim that mirroring is the proper way to describe the crucial brain activities in the experiments in question. Indeed, one could well challenge the application of the label “mirror neurons” to logically related neurons. Simply because they are motor neurons, and are found in a brain region rife with MNs strictly so called, it doesn’t follow that they are mirror neurons. Given their behavior as described by Iacoboni et al., they seem to lack the congruence properties required for mirror neurons. In defense of Iacoboni et al.’s claims, it might be possible to loosen the definition of a mirror neuron in a plausible way to allow logically related neurons to qualify. I won’t pursue this possibility.⁵

However, a case can still be made for mirror-based intention attribution as long as the relevant region (inferior frontal cortex) is the site of some mirroring activity, and as long as that mirroring activity makes causal contributions to intention attributions (whatever the substrates of those attributions). It is probable -- at least extremely plausible -- that classical mirror neurons as well as logically related neurons play causal roles in the entire process. When a participant in the Iacoboni et al. study observes the Intention video clip and sees a hand grasping a cup, presumably his classical MNs for grasping respond in a resonating, or mirroring, fashion. This mirroring activity by classical MNs may well generate the (non-mirroring) activity in the logically related neurons, which in turn leads to an intention attribution (whatever the substrate of that attribution). If all of this is correct, it is a case in which mirroring makes some causal contribution to intention attribution. Jacob would want to claim that this isn't the kind of intention attribution Gallese and Goldman (1998) had in mind because it doesn't display the retrodictive pattern. I shall return to this point below. But even if it isn't retrodictive, it would still be mirror-caused mindreading, and would therefore support the MM thesis.

It is worth emphasizing that the Iacoboni et al. paper presented evidence for attributions of a prior intention that doesn't depend on an arguably controversial interpretation of neural activations revealed by functional imaging. 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" conditions. 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

there is independent evidence of intention attribution, where the attribution was often spontaneous. Thus, a highly probable scenario is that observers first mirrored the agent's grasping intention, which led (together with observation of the context) to a replication of the intention to drink (or clean up), and finally to an attribution of the latter (so-called "prior") intention. This would qualify as mirror-caused mindreading under my definition.

Jacob makes a third point in this neighborhood, one that isn't aimed at the Fogassi et al. and Iacoboni et al. studies per se, but at a putative difficulty they pose for the original conjecture of mirroring and mindreading that Gallese and Goldman proposed in 1998. Gallese and Goldman conjectured that MNs play a role in the retrodictive reading of an agent's mental state, whereas the Fogassi et al. and Iacoboni et al. studies posit a predictive reading of an agent's (future) mental state. Jacob argues as follows:

[O]ne cannot have it both ways: either MN activity is predictive (in accordance with the new model of logically related MNs) or it is retrodictive (in accordance with Gallese and Goldman's conjecture). In accordance with the new model, it is, I think, more plausible to choose the former option, on which MN activity is predictive. (Jacob, 2008, PDF, p. 35)

Is there really a "tension" between the evidence from the Iacoboni et al. study and the original Gallese-Goldman conjecture?

The first point to notice is that when a subject concludes from the Intention video clip that the agent has a ("prior") intention to drink or to clean up, it isn't strictly correct to describe this as a future intention. He isn't engaged in predicting an intention that will occur later. Instead, he is inferring a concurrent intention to perform an action in the

future, which is not predicting a future intention (i.e. an intention that will occur in the future). In fact, the observer's thinking is as much retrodictive as in the ordinary case of inferring an intention to perform a currently observed action (of the kind that concerned Gallese and Goldman).

Suppose, however, that an experiment provided genuine evidence of mirroring being involved in predictions of intentions. Would this pose a problem? Jacob argues that MN activity is either predictive or retrodictive, where the 'or' is exclusive. "[O]ne cannot have it both ways," he says. Why should this claim be accepted? Why couldn't mirroring sometimes play a (causal) role in retrodictive tasks of intention attribution and sometimes play a (causal) role in predictive tasks of intention attribution? These might be entirely separate types of cases, but mirroring might play a role in both. So this "dilemma" doesn't really get off the ground. The principal problem is the assumption that MNs have only a single function, predicting or retrodicting, an assumption that isn't defended and seems to be unmotivated.⁶

III

I conclude that the evidence adduced for a mirroring-mindreading link in the motor domain is quite suggestive, though not fully probative. Even if one takes a more critical stance toward the intention prediction studies, this would not be a serious setback for the original MM thesis that was presented in section I. MM is a thesis about mirroring and mindreading in general, not a thesis about motor mirroring in particular. In saying that mirroring processes play an important role in mindreading, no premium is placed on motor mirroring specifically. It is natural to assume that evidence for mirror-

based mindreading should come first and foremost from the motoric domain, because motor mirroring was the first type of mirroring discovered and the first that was hypothetically linked to mindreading. These are historical accidents, however. Mirror processes outside the motoric domain are now well established, especially for emotion and sensation. There is no intrinsic reason why motor mirror processes should take pride of place. If there is good evidence for mirror-based mindreading in these other domains, that should suffice as evidence for MM. This is so especially if we don't make exaggerated claims for the role of mirroring in mindreading. I shall say more about the scope of mirror-based mindreading in section IV. At this juncture, I am interpreting MM in a weak sense, as simply saying that there exists some mirror-based mindreading. In the rest of this section, I present evidence for this kind of mindreading.

There is now substantial evidence, I submit, for mirror-based mindreading in several domains, including evidence about emotion and sensation. Although Jacob acknowledges the existence of such evidence (see his footnotes 7 and 33), he decides to set it aside in his target paper. The unfortunate result is, first, that a reader may get an incomplete and imbalanced picture of the current state of the scientific evidence for the MM thesis. Second, and of special relevance to the current response, since Jacob includes my work as part of his critique and specifically discusses several points arising in *Simulating Minds*, a reader might infer that Jacob has covered the major evidence for the MM thesis I adduce in that book. In fact, the most important and extensive chunks of evidence presented there make no appearance in his discussion.⁷

Evidence for mirror-based mindreading in the domain of emotion can be assembled from two types of sources (Goldman, 2006; Goldman and Sripada, 2005).

First, one needs evidence of a mirror process for one or more emotions, which can be established by fMRI studies. Second, one needs evidence that the mirror process in question is causally implicated in attributions of the emotion in question, at least in a certain kinds of conditions, e.g. observing someone's facial expression. Evidence for this causal thesis, as it happens, is best extracted (inferentially) from neuropsychological evidence.

To probe the possibility of mirroring for disgust, Wicker et al. (2003) scanned normal participants both during their own experiences of disgust and during observation of other people's disgust-expressive faces. The participants were scanned while viewing movies of individuals smelling the contents of a glass (disgusting, pleasant, or neutral) and forming spontaneous facial expressions. The same participants were also scanned while inhaling disgusting or pleasant odorants through a mask. The finding was that the same areas, the left anterior insula and the right anterior cingulate cortex, were preferentially activated both during the experience evoked by disgusting odorants and during observation of other people's disgust-expressive faces. This establishes a mirroring process for disgust. However, Wicker et al. didn't have participants perform any emotion recognition tasks, so they didn't obtain any information about the attribution of disgust, or any possible link between the mirroring of disgust and disgust attribution.

There are lesion studies, however, that present evidence highly relevant to this question. Here is one such item of evidence. Calder et al. (2000) studied patient NK who suffered damage to the insula and basal ganglia. In questionnaire responses NK showed himself to be selectively impaired in experiencing disgust. Other tests showed that he was also significantly and selectively impaired in disgust recognition, i.e., attribution.

His selective impairment in disgust attribution is best explained by the selective damage to his experience of disgust. Because he couldn't experience disgust normally, he didn't experience it when viewing others' disgust-expressive faces, as normal people would do whose disgust-mirroring capacity is intact. This strongly suggests that in the normal process of attributing disgust to another (based on an observed facial expressions), a mirrored, resonant experience of disgust is a crucial causal link. (For detailed arguments against rival explanations of the selective impairment in attribution, see Goldman and Sripada, 2005; Goldman, 2006: 117-132.) An analogous pattern of findings also exists for the case of fear, but there are complications in the fear story that shouldn't detain us here.

In the domain of sensation, ample evidence supports both the mirroring of pain and the thesis that such mirroring plays a causal role in third-person pain attribution. Singer et al. (2004), Jackson et al. (2004) and Morrison et al. (2004) all reported findings of pain mirroring or resonance. Those reports were restricted to the affective portion of the pain system, but subsequent studies by Avenanti et al. (2005, 2006), using TMS, found empathy for pain in the sensorimotor part of the pain system. On the question of whether mirrored pain can play a causal role in pain attribution, results from both Jackson et al. and from Avenanti et al. are affirmative. Jackson et al. had subjects watch depictions of hands and feet in painful or neutral conditions and were asked to rate the pain intensity they thought the target was feeling. Such a rating is a third-person attribution task. There was a strong correlation between the ratings of pain intensity and the level of activity within the affective portion of the attributors' own pain network. Avenanti et al. (2005) analyzed subjective judgments about the sensory and affective

qualities of the pain ascribed to a model while watching a video in which a sharp needle was shown being pushed into the model's hand. These judgments of sensory pain in the model seemed to be based on the mirroring process in the attributor's own sensorimotor pain system (see Goldman, 2008a, for a brief review).

There are other confirmed examples of mirroring in the sensation domain, including mirroring for touch. In a fMRI study, 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. Blakemore et al. (2005) provide dramatic support for the mirroring of touch, showing that mirroring events can rise above the threshold of consciousness. However, there haven't been tests to determine if the observation of someone else being touched causes judgments to the effect that the model undergoes specified feelings of touch. But the absence of positive evidence isn't negative evidence, and previous discoveries in analogous domains lead one to expect that positive evidence could be obtained if the relevant studies were done. Returning to the general picture, we can say at the present time that MM is positively supported in selected domains of emotion and sensation. It is unfortunate that Jacob gives scant attention to these findings.

IV

A final element of Jacob's critique I shall discuss concerns the relationship between mirroring and the simulation theory (ST) of mindreading. Gallese and Goldman (1998) defended ST by presenting mirroring as the medium of mental simulation. But is all mental simulation mirroring? Is any of it mirroring? Jacob has his doubts.

Could MNs constitute a primitive version of a mental simulation heuristic, as Gallese and Goldman conjecture? Since most early ST approaches to mindreading appealed to the concept of pretence, one obvious challenge for Gallese and Goldman's ambitious research program is to show that both pretence and interpersonal mirroring exemplified by MNs are instances of mental simulation. This is the challenge to which I turn in the present section. These are two different strategies for implementing Gallese and Goldman's (1998) research program and, as I shall shortly argue, they part company on how exactly to fill the details of this program. (Jacob, 2008: p. 12 of PDF)

As Jacob reports, Gallese and collaborators develop a notion of "embodied simulation," which is sometimes linked to so-called internal models of action invoked by proponents of control theory. I shall not explore this proposal in any depth. I confine myself to the remark that reliance of the internal-model approach on notions like efference copy seems to restrict simulation to the motoric domain and isn't readily extendable to simulation or mirroring processes for emotion or pain. It seems doubtful, then, that this approach can support an appropriately inclusive simulation theory of mindreading (see Goldman, 2008b).

I now focus attention, however, on Jacob's doubts about my own favored approach. On the view developed in *Simulating Minds* (Goldman, 2006), mirroring is not the only form of mental simulation, and mirror-based mindreading is not the only form of simulation-based mindreading. Jacob is right that "pretense" was the core concept in the original simulation approach (Gordon, 1986; Heal, 1986; Goldman, 1989; Currie, 1998), and pretense should not be identified with mirroring. In *Simulating Minds* I first develop

a generic concept of simulation as a process that either resembles or duplicates a target process in relevant respects or tries to resemble or duplicate it in relevant respects.⁸ Mental simulation is simulation in which one mental process matches another one, or is launched in an attempt to match another one. Mental simulation can be either intra-personal or inter-personal; in third-person mindreading, it is obviously inter-personal. I then argue that there are two forms of mental simulation. The first is mirroring, which is automatic, almost entirely unconscious, typically involves comparatively “primitive” mental states, and doesn’t rely on task-specific knowledge or information. I call this kind of simulation “low-level simulation.” This is to be distinguished from “high-level simulation,” which corresponds to the original idea of pretense-driven simulation, or, in the language of *Simulating Minds*, imagination-driven simulation. This form of simulation is more effortful, is sometimes conscious, characteristically involves more complex mental states, and is guided by task-specific knowledge or information. High-level simulation is a process that aims to replicate another mental state or mental process, but unlike the case of mirroring, there is no mechanism or pre-packaged process that normally guarantees success (i.e. genuine matching or resemblance). Success depends heavily on (inter alia) the quality of the stored information that guides the simulation.⁹

The next question is whether any mindreading is based on low-level simulation and whether any mindreading is based on high-level simulation. I defend affirmative answers in both cases. We reviewed evidence for the low-level case in section III. Evidence for the high-level case is much less straightforward and I won’t try to review it (see Goldman, 2006: chaps. 7-8). However, in comparing the position of *Simulating Minds* to the proposal of the 1998 Gallese-Goldman paper, one non-trivial change should

be noted (perhaps anticipated by Jacob). Although the Gallese-Goldman paper proposed to ground traditional ST in mirroring processes, it now appears that mirror-based mindreading occurs (at most) in a class of cases rarely discussed by traditional ST (or TT). These are cases such as motor intention attribution, face-based emotion attribution, pain reading, and the like. The standard examples of mindreading that figured in the traditional debate, by contrast, were decision attribution, desire attribution, belief attribution, and the like. According to *Simulating Minds*, none of the latter are to be explained by mirroring. This is not an objection to the approach, however. Low-level mindreading exists and needs to be explained, despite being largely ignored in the early literature. The mirroring variant of simulation appears to be a promising approach that provides the needed explanations.

In *Simulating Minds* all my examples of low-level mindreading are explained in terms of mirroring. Do I mean to hold that mirroring is a necessary condition for low-level mindreading (or for mindreading in general¹⁰)? I don't mean to make this a definitionally necessary condition for low-level mindreading, and I don't believe that the definitions I offer of mirroring and of low-levelness have this implication. Nor do I believe that there is a psychological "law" that somehow guarantees that all low-level mindreading involves mirroring (a question Jacob poses to me in a personal communication). So, I am open to the possibility of other forms of low-level mindreading that don't display mirroring. Jacob, however, proceeding under the assumption that I am committed to the thesis that all low-level mindreading involves mirroring, offers a class of cases he regards as a counterexample.

Since the work of Heider and Simmel (1944), it is known that subjects ascribe goals, emotions, and social intentions to moving geometrical stimuli. Perhaps even infants or toddlers have some such trait, because even 12-month-olds exhibit a preference for ‘helping’ motions over ‘hindering’ motions as executed by geometrical stimuli (Kuhlmeier, 2003). Jacob concludes that “it is highly unlikely that human infants represent the intentions of moving geometrical stimuli by a process of motor simulation of (or motor resonance with) the latter’s non-biological motion.” So there is low-level mindreading without mirroring, Jacob apparently concludes.

First, I don’t think we can confidently say that the responses of 12-month-old toddlers reflect mindreading. A “preference” for helping versus hindering does not clearly attest to mindreading. Second, even if we concede that these are cases of mindreading, and presume that they are low-level mindreading, is it clear that they don’t involve mirroring? What precludes mirroring in these cases? There is the obvious fact that geometrical stimuli don’t themselves have brains that experience goals, emotions, or social intentions, so the observers’ own goals, emotions, or social intentions (if these states are tokened) could not match or replicate anything in the targets. But this is only a problem if ‘mirroring’ is understood as core mirroring, as defined above in section I. Under the second definition, however, the definition of generic mirroring, this isn’t a problem. What I suspect Jacob thinks is that all mirroring is motor mirroring and this only occurs when observing biological motion, which is not what the geometrical stimuli display. However, as Jacob himself reports (footnote 6), Romani et al. (2005) report motor facilitation prompted by the observation of biologically impossible finger movements. More importantly, there can be mirroring that involves no motor

responsiveness. As shown in section III, studies of disgust, touch, and pain reveal processes that satisfy a plausible definition of mirroring. So, even if there is no covert motoric mimicking by observers of moving geometrical stimuli, this doesn't preclude a mirroring process. Nonetheless, it wouldn't affect any important commitment of mine to discover that there is low-level mindreading without mirroring.

I turn now to an additional problem Jacob raises for my overall simulationist approach.¹¹ Toward the end of section 3.1 of his paper, he airs a worry about the interface between the mindreading of motor intentions (intentions to perform something like "basic" actions) and the mindreading of higher-level intentions. The worry seems to be this: on the assumption that motor intentions are mindread by mirroring mechanisms and that higher-level (or "prior") intentions are mindread otherwise, how is it possible to "bridge the gap" between them?

Motor resonance on its own lacks the resource to bridge the gap between representing an agent's motor intention and representing her higher-level intentions. The question is: when motor resonance is indeed used, what further cognitive resources will enable a mindreader to move from a non-conscious representation of the agent's motor intention to a conscious representation of her prior intention, and thereby form a belief about the content of that prior intention?

(Jacob, 2008: pp. 27-28 of PDF)

This worry seems to have two elements. One element concerns consciousness versus non-consciousness. The representation of the agent's motor intention is likely to be non-conscious whereas the representation of the agent's prior intention is likely to be conscious. How is that gap to be bridged? The second element concerns representations

with conceptual versus non-conceptual content. Earlier Jacob writes: “So an agent’s motor intention, unlike her prior intention, has non-conceptual content.” (2008, p. 24 of PDF) He apparently regards this too as a difficult-to-bridge gap.

I do not see why these questions are special problems for the simulation approach. That is particularly true in the case of the consciousness/non-consciousness “gap.” Vast stretches of cognitive activity involve combinations of conscious and non-conscious processing. Whether the topics are vision, memory, or mindreading, there are great swaths of non-conscious processing that pass information along to conscious processes. How exactly this is done is a question for everyone; nobody has a very good handle on it. I don’t see why a special onus rests on simulation theorists to answer it – to show how this type of “gap” is closed.

Mirroring or resonating is primarily a non-conscious process, but why should it be surprising that its outputs can be “passed along” to conscious awareness. As the Wicker et al. (2003) experiment shows, when an observer looks at a face that expresses disgust, mirroring activity is activated in the observer’s brain. Studies of face-based emotion recognition make it clear that a normal observer of such a face will classify the emotion as disgust. This act of classification – whether executed by a button press or some other method – is likely to be conscious. So, how does the non-conscious mirror processing produce a conscious classificatory act? This is not a well-studied topic, but why should it be an insuperable burden for a mirroring or simulation approach? Cognitive science generally lacks a firm understanding of the consciousness/unconsciousness interface, so this cannot be viewed as a special defect of mirror theories or simulation theories.

In fact, the study of mirroring may shed new light on the consciousness/non-consciousness divide. Blakemore et al. (2005) showed that mirroring events are capable of rising above the threshold of consciousness. They describe a subject **C** for whom the observation of another person being touched is experienced consciously 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. This illustrates the point that there is no unbridgeable "chasm" between mirroring and consciousness. In normal subjects, of course, mirroring events are unconscious, but why should there be an in-principle problem with their causally interacting with conscious events? True, we have no general answer to the question of why some cognitive events are conscious and others unconscious, but this is not the basis for a sound critique of either mirror theory or simulation theory.

The second element in Jacob's worry concerns the interaction between non-conceptual representations of motor intentions and conceptual representations of higher-level (or "prior") intentions. Jacob doesn't tell us exactly what he means by the conceptual/non-conceptual distinction, and different writers draw the distinction differently (see Byrne, 2005). Let's cede him the distinction, however, and ask where it might take us. The worry again seems to pertain to the prospects for "communication" between different types of mental representation, this time between conceptual versus non-conceptual mental representations. Again I counter by saying that this isn't a distinctive problem for simulation or mirroring theory. One such problem in cognitive science concerns different ways of representing space, a problem explored by Jackendoff (1996). There seem to be visual ways of encoding information about space (arguably,

non-conceptual representations) and also linguistic or conceptual ways of encoding information about space. But the mind enables these differently formatted representations to communicate with one another. How does this transpire? This is a good question to which there may not be known answers. The point is that this is a pretty ubiquitous problem, and whatever its solution, there is no obvious reason why simulation theory or mirroring theory cannot avail themselves of it. Encountering a challenge common to many quarters of cognitive science does not seem like a serious objection.

References

- Avenanti, A., Buetti, 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.
- 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.
- Byrne, A. 2005: Perception and conceptual content. In M. Steup and E. Sosa, eds., *Contemporary Debates in Epistemology*. Malden, MA: Blackwell Publishing.
- Calder, A.J., Keane, J., Manes, F., Antoun, N. and Young, A.W. 2000: Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, 3, 1077-1078.
- Currie, G. 1998: Pretence, pretending and metarepresentation. *Mind & Language*, 13, 35-55.
- 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.
- Gallese, V. and Goldman, A. I. 1998: Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2, 12, 493-5-1.

- Gallese, V., Keysers, C. and Rizzolatti, G. 2004: A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8, 9, 396-403.
- Goldman, A. I. 1989: Interpretation psychologized. *Mind & 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. 2008a: Mirroring, mindreading, and simulation. In J. Pineda, ed., *Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition*. Humana Press.
- Goldman, A. I. 2008b: Does one size fit all? Hurley on shared circuits. *Behavioral and Brain Sciences*.
- Goldman, A. I. 2008c. Two routes to empathy: Insights from cognitive neuroscience. In A. Coplan and P. Goldie, eds., *Empathy: Philosophical and Psychological Perspectives*. New York: Oxford University Press.
- Goldman, A. I. and Sripada, C. S. 2005: Simulationist models of face-based emotion recognition. *Cognition*, 94, 193-213.
- Gordon, R. M. 1986: Folk psychology as simulation. *Mind & Language*, 1, 158-171.
- Heal, J. 1986: Replication and functionalism. In J. Butterfield, ed., *Language, Mind, and Logic*. Cambridge: Cambridge University 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, 3, 529-535.
- Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J. and Rizzolatti, G. 1999: Cortical mechanisms of human imitation. *Science*, 286, 2526-2528.

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.

Jacob, P. 2008: What do mirror neurons contribute to human social cognition? *Mind & Language*.

Jacob, P. Under submission: What if mirroring turns out to be a by-product?

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.

Morrison, I., Lloyd, D., di Pellegrino, G. and Roberts, N. 1992: Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue? *Cognitive, Affective, Behavioral Neuroscience*, 4, 270-278.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. and Rizzolatti, G. 1992: Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91, 176-180.

Romani, M., Cesari, P., Urgesi, C. Facchini, S. and Aglitoi, S. M. 2005: Motor facilitation of the human cortico-spinal system during observation of bio-mechanically impossible movements. *NeuroImage* 26, 755-763.

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

de Vignemont, F. and Haggard, P. In press: Action observation and execution: What is shared? *Social Neuroscience*.

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 Notice that the constitution relation, as understood here, does not have the status of a definition or some other conceptual relationship. It is a matter for empirical discovery. That Michelangelo's statue of David is composed of, or constituted by, a lump of marble is a matter to be determined empirically rather than by definition. Similarly, which neural events constitute an act of mental-state attribution -- which ones are the substrate of such an act -- is something that must be determined empirically.

² In footnote 2 of his paper, Jacob indicates that he uses "rehearse" interchangeably with "resonate", "match" and "replicate." So his phrase "mental rehearsal" refers to mirroring activity per se, not something merely caused by mirroring activity.

³ Another worry (but one I won't press here) is that a mirroring event would not have as part of its content a reference to the agent. If the mirroring event that occurs in the observer's brain is genuinely congruent with an event in the agent being mirrored, it must use a first-person indexical in referring to the actor whose motor act is being planned or commanded. (A first-person indexical would appear as part of the "character" of the thought, in David Kaplan's sense.) However, if this is the nature of the observer's thought, it cannot refer to the target agent to whom the observer ascribes an intention. Hence, it must be some other cognition, not the mirroring cognition per se, that constitutes an intention attribution to the target agent. For further discussion of the problem of congruency, or shared representations, see de Vignemont and Haggard (in press).

⁴ The label "retrodictive" attribution is slightly unfortunate, because the attributed intention is only very slightly in the past. It might better be called "concurrent" attribution. It was called "retrodictive" because the ascribed motor plan is a presumed cause of the observed behavior and hence prior to it.

⁵ It might be thought that my own definition (M) of a mirroring process is already loose enough for these purposes, because (M), unlike definition (M_c), does not require actual congruence. But this isn't correct, as far as I can see. Definition (M) specifies that a mirror process is one in which the normal situation would involve congruence, and this requirement does not seem to be satisfied for logically-related neurons.

⁶ Even if we accepted the "one function only" thesis, how should functions be individuated? Is mindreading a single function? Or is mindreading beliefs one function, mindreading intentions another function, etc.? Unless it is clear -- which it isn't -- what counts as a single function, how can one say that retrodictive mindreading is one function and predictive mindreading a second function?

⁷ Jacob discusses the possibility of non-motoric mirroring -- on which he renders a negative verdict -- in a separate paper (Jacob, under submission). But this is not the occasion to examine his not-yet-published arguments.

⁸ This isn't the full definition, but other details can be neglected for present purposes.

⁹ In a forthcoming paper (Goldman, 2008c) I make reference to a neural network that might be a good candidate for a substrate of *high-level* mindreading. No such substrate was proposed in *Simulating Minds*, however.

¹⁰ In his footnote 10, Jacob tries to show that I am committed to the principle that mirroring is a necessary condition of mindreading, given other things he takes me to hold. It is clear from what I have said above, however, as well as from *Simulating Minds*, that I don't hold this view, first, because I don't treat high-level mindreading in terms of mirroring, and second, because I don't hold that all mindreading is simulational (see note 11 below.) Also, Jacob's alleged deduction of this view from my purported premises is quite odd. He says that from the two premises (i) that mirroring is one instance of mental simulation and (ii) that mental simulation is a basis for mindreading it follows (iii) that mirroring is a necessary condition

for mindreading. Conclusion (iii) does not follow from (i) and (ii), at least if we understand (iii) to say that all acts of mindreading involve mirroring.

¹¹ It should be added that *Simulating Minds* does not claim that all mindreading is executed by simulation. It leaves room for theory-based mindreading, and hence the overall approach is a simulation-theory hybrid. The emphasis, however, is on the simulation component, and the present discussion is restricted to that component, the only one Jacob addresses.