If it is true that whatever state (2) is, someone in (3) can know that Mr Nicholls is in that state (that is, that the description of the state can be as particular as it needs to be), then it is not clear that feeling the emotion someone feels is epistemologically better than knowing that they are feeling that emotion. It does seem that, for any state, there is a proposition that will describe that state. Even if the particularity of an emotion exceeds the capacities of a brief statement in English, that does not mean that it is beyond the capacities of a longer statement, or perhaps a statement in another language (Budd (1995): 146).[5] One might nonetheless think that (2) would teach the reader *what Mr Nicholls' experience was like*. However, we are assuming that the reader is familiar with the types of emotion Mr Nicholls is undergoing, so that would not be something he or she could learn. The second issue seems to me more promising for a defence of the epistemological status of empathy. As the causal capacities of a mental state depend on more than its propositional contents, (2) will be a better replicator of the causal role of (1) than will (3). Believing that Mr Nicholls feels anxious might leave the reader in the dark about what he feels inclined to do, whilst feeling his anxiety (or something like his anxiety) might provoke the intense off-line desire to bolt out of the room, or some complex combination of desires that would include that desire vying with the desire to grab Charlotte Bronte. If this is true, then someone in (2) will be in a better position to judge Mr Nicholls' position than someone only in (3). In this respect, then, empathy (if it is possible) has an advantage over its purely cognitive rival.

[5] Jane Heal presents powerful reasons for doubting this claim Heal (1997/2003a).

# 3

# Two Routes to Empathy:

## Insights from Cognitive Neuroscience

*Alvin I. Goldman*

## 3.1 Definitional Overview

The concept of empathy has a considerable history in both philosophy and psychology, and may currently be enjoying an apex of attention in both. It is certainly receiving close attention in cognitive neuroscience, which brings fresh discoveries and perspectives to the subject. The term 'empathy', however, does not mean the same thing in every mouth. Nor does there seem to be a single, unified phenomenon that uniquely deserves the label. Instead, numerous empathy notions or phenomena prance about in the same corral, and part of the present task is to tease some of these notions apart. More importantly, there are fascinating new findings that should be reported, analyzed, and mutually integrated, whether one's interest in empathy is primarily driven by pure science, philosophy of mind, moral philosophy, or aesthetic theory.

As a first step in distinguishing multiple senses, grades, or varieties of empathy, consider a definition offered by Vignemont & Singer (2006):

> There is empathy if: (i) one is in an affective state; (ii) this state is isomorphic to another person's affective state; (iii) this state is elicited by the observation or imagination of another person's affective state; (iv) one knows that the other person is the source of one's own affective state. (2006: 435)

Questions can be asked about this definition that might motivate alternative definitions. For example, clause (i) restricts empathic states to affective or emotional states, but this is too narrow for some purposes. Cognitive neuroscientists talk of empathy for touch (Keysers et al. (2004)) and empathy for pain (Singer et al. (2004); Jackson et al. (2004); Morrison et al. (2004)), but neither touch nor pain is usually considered an emotion (although pain has an affective dimension as well as a sensory one). Concerning clause (ii), it should be asked exactly what is meant by 'isomorphic'. If it means a state of one person that matches a state of the target, then that requirement is more restrictive than definitions offered by others. Hoffman (2000), for example, defines empathy as 'an affective response more appropriate to another's situation than one's own'. This doesn't imply that the receiver's affective state matches (or is isomorphic to) that of the target.

Clause (iii) might be questioned on a rather different ground. It seems right to restrict empathic states to ones acquired by observation or imagination of the target individual. But shouldn't the elicitation process be constrained even further? For example, David Hume writes:

> 'Tis indeed evident, that when we sympathize with the passions and sentiments of others, these movements appear at first in *our* mind as mere ideas, and are conceiv'd to belong to another person, as we conceive any other matter of fact... No passion of another discovers itself immediately to the mind. We are only sensible to its causes or effects. From *these* we infer the passion: and consequently *these* give rise to our sympathy. (1739–1740 (1978): 319, 576)

Hume (using the term 'sympathy' rather than 'empathy') apparently endorses a three-stage hypothesis: one observes another person's movements, one infers from those movements a certain passion in the person, and the inferred belief causes a matching passion in oneself. If this is right, the process satisfies the Vignemont-Singer definition because the affect is elicited—albeit indirectly—by observation. But many people conceptualize empathy as a spontaneous, non-inferential process. If they wish to define empathy in that fashion, the previous definition would have to be amended to exclude inferential steps.

Another dimension of empathy important to many theorists is 'care' or 'concern' for the target. This dimension is omitted in the Vignemont-Singer definition. Social psychologists are traditionally interested in empathy as the basis of altruistic behavior, and many would want to highlight that component of empathy. Other investigators are interested in empathy as a key to mindreading, and might even use the term 'empathy' to describe (what they take to be) the most common form of mindreading. In other words, they use the term 'empathize' as roughly equivalent to 'simulate' (in an intersubjective fashion). I myself am a partisan of this position (Goldman 2006a), but this will play only a secondary role in the present paper. The proffered definition is neutral on the question of mindreading, and that's fine for present purposes.

It is easy to conflate different features of empathy, so readers can sometimes be mystified as to how, exactly, a given writer uses the term. For example, in Baron-Cohen's (2003) account of autism, or Asperger's syndrome, the linchpin of the account is a deficiency in 'empathizing'. But in reading Baron-Cohen it is often difficult to tell which of three possible senses of 'empathizing' he primarily has in mind: (A) using simulation when engaging in mentalizing, (B) being curious about others' mental states, or (C) feeling concern about other people's feelings. Correspondingly, a deficiency in empathizing might consist in a sparse use of simulation, a dearth of curiosity about others' mental states, or a low-level of concern about other people's feelings.

These preliminary comments should alert the reader to the fact that different writers and researchers exhibit different approaches to empathy. In addition, however, research findings can contribute to an understanding of how empathy is produced. Is there exactly one route to empathy, that is, one cognitive system—or one *type* of cognitive system—that produces empathy, or is there more than one? How exactly does this system, or these systems, work? What different consequences might ensue as upshots of different modes of empathizing? These are the primary questions to which this paper is addressed.

## 3.2 The Mirroring Route to Empathy

In an earlier era, one might have been skeptical about the isomorphism, or matching, condition we provisionally accepted in the definition of empathy. Do empathizers really undergo states that match those of their targets? Are the feeling states of receivers exactly the same as those of their targets? Since the discovery of mirror neurons and mirroring processes, however, there is much less room for skepticism. There is little doubt about the existence of processes through which patterns of neural activation in one individual lead, via their observed manifestations (e.g. behavior or facial expressions), to matching patterns of activation in another individual. If the corresponding patterns of activation are not perfect duplicates, at least they resemble their corresponding states in the target in terms of the kinds or types of mental or brain activity involved. Some might balk at calling the resonant states 'mental' states, because the mirroring episodes commonly occur below the threshold of consciousness even when the episodes being mirrored are fully conscious. If the term 'mental' is used broadly, however, they are processes of 'mental mimicry.'

Mirror neurons and mirroring processes were first discovered in monkeys, and subsequently in humans, in connection with preparation for motor action (Rizzolatti et al. (1996); Gallese et al. (1996)). When a monkey plans a certain type of goal-related hand action, e.g. tearing, holding, or grasping, neural cells in its premotor cortex dedicated to the chosen type of action are activated. Surprisingly, when a monkey merely observes another monkey or human perform a similar hand action, the same cells coded for that type of action are also selectively activated. Thus for certain neurons there is a sort of neural mirroring; one thing that occurs in the actor's brain is (more or less) replicated in the brain of the observer. These kinds of cells were therefore dubbed 'mirror neurons.' There are many details concerning the precise activation properties of mirror neurons in an observer versus an actor (Rizzolatti et al. (2001)). But the basic finding is that there is robust, selective activation of the same cells in both execution and observation modes.

Using different techniques, an action-related mirror system has been found in humans, centered on the inferior parietal lobule and the premotor cortex, including Brodmann area 44 (see Rizzolatti & Sinigaglia (2008)). Cochin et al. (1998) showed that the same $\mu$ rhythm that is blocked or desynchronized when a human performs a leg or finger movement is also blocked when he merely observes a similar movement by another person. Similar results were obtained from research studies using magnetoencephalography (MEG) and transcranial magnetic stimulation (TMS). Fadiga et al. (1995) recorded the motor evoked potentials (MEPs) induced by magnetic stimulation of the left motor cortex in various muscles of the contralateral (right) hands and arms of subjects who were watching the experimenter either grasp objects with his hand or make movements unrelated to any object. In both cases a selective increase in MEPs was found in the recorded muscles. Thus, mirroring properties were detected both for the observation of goal-related actions, as in monkeys, and also for non-object-related arm movements, which is not found in monkeys.

A study by Buccino et al. (2001) showed that mirroring for action isn't restricted to actions of the hand or arm. Subjects were shown action stimuli of the following sorts: biting an apple, grasping a cup, kicking a ball, and non-object-related actions involving the mouth, hand, or foot. The results showed that observing both object-related and non-object-related actions led to the somatotopic activation of the premotor cortex, with the mouth represented laterally and the foot medially.

Which mental states are activated in the case of motor mirroring? As I have said, it is presumably plans or intentions to do specific actions. Matching motor plans are activated in the observer, but they don't normally lead to imitation. Their outputs are usually inhibited downstream. There is mental mimicry, one might say, but not behavioral mimicry.

Mimicry of action-planning states doesn't naturally invite the label of empathy. But many other mental states that partake of mirroring more naturally invite talk of empathy. Some writers might prefer other labels. One might speak of 'resonance,' for example, or 'contagion.' But I think that 'empathy' is a reasonable choice. It must be stressed, however, that in many mirroring activities the receiving end of the mirroring relationship may not be conscious. The receiver may not be aware, or not fully aware, of the mental event she is undergoing that happens to be congruent with an event in the sender. This may raise issues concerning condition (iv) of the definition discussed earlier. I think it is fair to require a receiver to have some sort of intentional attitude directed toward the target by which the resonating state is linked to him. Otherwise, it doesn't seem like a case of empathy. I suspect that condition (iv) is too strong an intentional condition of this kind, but I don't have a wholly suitable replacement for it.

Even if a suitable replacement for condition (iv) is found, 'empathy' might not be a tempting term for mental mimicry of action-planning. Let us therefore examine other categories, starting with the sensation of touch. Keysers et al. (2004) found that when a person watches another person being touched, the same brain areas are activated as those in the person being touched. More specifically, they found that touching a subject's own legs activated the primary and secondary somatosensory cortex of the subject. Large extents of the secondary somatosensory cortex also responded to the sight of someone else's legs being touched. Films used with control subjects in which the same legs were approached by an object, but never touched, produced much smaller activations. This phenomenon is naturally described as empathy for touch.

Another mirroring domain involves the sensation of pain. Pain is a complex sensory and emotional mental state associated with actual or potential body damage. Sensory components of pain evaluate the locus, duration, and intensity of a pain stimulus, and affective components evaluate the unpleasantness of the noxious stimulus. These are mapped in different nodes of the so-called 'pain matrix.' Sensorimotor cortices process sensory features of pain and display somatotopical organization (mapping locations of the stimuli in brain tissue). Affective and motivational components of pain are coded in the affective node of the pain matrix, which includes anterior cingulate cortex (ACC) and anterior insula (AI). The subjective feeling of unpleasantness is strictly associated with neural activity in these structures.

In 2004 mirroring for pain was established in three articles: Singer et al. (2004), Jackson et al. (2004), and Morrison et al. (2004). In each of these studies empathy for pain elicited neural activity mainly in the affective division of the pain matrix, suggesting that only emotional components of pain are shared between self and other. However, using transcranial magnetic stimulation, Avenanti et al. (2005, 2006) found that the direct observation of painful stimulations on a model elicits inhibitory responses in the observer's corticospinal motor system similar to responses found in subjects who actually experience painful stimulations. When participants watched a video showing a sharp needle being pushed into someone's hand, there was a reduction in corticospinal excitability in related muscles. No change in excitability occurred when they saw a Q-tip pressing the hand or a needle being pushed into a tomato. These 'mirror' responses were specific to the body part that the subjects observed being stimulated and correlated with the intensity of the pain ascribed to the model, thus hinting at the sensorimotor side of empathy for pain.

The best example of mirroring in the sphere of emotions features the emotion of disgust, and the clearest evidence comes from an fMRI study by Wicker et al. (2003). Participants were scanned while passively inhaling disgusting or pleasant odorants through a mask and, separately, while observing movies of individuals who smelled the contents of a glass (disgusting, pleasant, or neutral) and spontaneously manifested appropriate facial expressions. The core finding was that the left anterior insula—previously known to be implicated in disgust experience—and the right anterior cingulate cortex were preferentially activated both during the inhaling of disgusting odorants (compared with pleasant and neutral odors) and during the observation of disgust facial expressions (compared with pleasure-expressive and neutral faces). This shows that observing a disgust-expressive face produces mental mimicry, or empathy, in an observer of the model. To use another expression very common in the literature, part of the observer's brain *simulates* the activity of a corresponding part of the model's brain.

In addition to the fMRI demonstration of matching experiences in observers and models in the gustatory cortex, researchers have used another measure of empathy to test whether observers experienced empathy. Jabbi et al. (2007) examined whether the IFO (anterior insula and adjacent frontal operculum) was associated with observers' self-reported empathy, measured by the Interpersonal Reactivity Index (IRI). They found that participant observers' empathy scores were predictive of their gustatory IFO activation while witnessing both the pleased and the disgusted facial expressions of others.

As is evident from the foregoing, a variety of systems in the human brain have mirror properties. They do not all use the same neural network or hardware. In particular, the mirror systems associated with sensations and emotions do not use the same neural hardware as the motor mirror system, nor as one another. Nonetheless, I shall treat them all as similar for present purposes, similar in having significant mirror properties. Ascending to an appropriate level of abstraction, we can consider them all to instantiate a single *type* of route to empathy, namely, a mirroring route. This does not imply that they all employ the very same cytoarchitectural pathway.[1]

## 3.3 A Reconstructive Route to Empathy

Granted that mirroring constitutes *one* (type of) route to empathy, is it the only type? This section presents two reasons to suspect otherwise. Mirroring seems to be, at least in one respect, automatic. The nature and content of mirroring events seem to be 'prepackaged'; they are not constructed on the fly. The disgust system, for example, is ready to respond to appropriate facial stimuli in a disgust-production mode. It doesn't have to manufacture a novel response to simulate the corresponding disgust experience in a model. Similarly, the action repertoire susceptible of motor mirroring is presumably pretty well fixed early in life. Although there is no consensus about the origins of mirror neurons, one promising hypothesis posits the work of the associative mechanism of 'Hebbian learning' (Heyes (2005); Keysers & Perrett (2004); Keysers & Gazzola (2006)). According to this hypothesis, mirror properties of visuomotor neurons are shaped in infancy, as a result of synchronous firing, and their subsequent activation should not require substantial online construction. In contrast with this automaticity of mirror-based empathy, a large chunk of empathy seems to involve a more effortful or constructive process. When empathizing with another, you often reflect on that person's situation, construct in imagination how things are (were, or will be) playing out for him, and imagine how you would feel if you were in his shoes. This process of perspective taking is the stuff of which most conscious empathizing, at any rate, is made. It doesn't have the effortless, automatic quality of mirroring (if mirroring is describable as having any 'quality' at all, i.e. any phenomenological 'feel'). This suggests that there is, indeed, a different kind of empathy in addition to mirroring. In fact, this other kind of empathy is more detectable in daily life than the mirroring kind, since

[1] However, it is also not implied that the various routes to empathy are non-overlapping. On the contrary, certain neural centers seem to be involved in the mirroring routes for several different sensations and/or emotions. Anterior insula appears to play a particularly important role in the processing of several such mental states (Singer et al., 2009).

mirroring is largely inaccessible to introspective awareness. Such a distinction between two types of empathy is embraced by Stueber (2006), who calls them 'basic' and 're-enactive' empathy respectively.

One must be careful in presenting the foregoing argument because some findings indicate that mirroring is not automatic in all respects. Singer et al. ((2006) and Vignemont & Singer (2006)) found that empathic responses to pain are modulated by learned preferences, and hence not purely automatic. In the Singer et al. experiment, participants played a Prisoner's Dilemma game in which confederates of the experimenters played either fairly or unfairly. Participants then underwent functional imaging while observing the confederates receive pain stimuli. The mirroring responses of the male participants were of special interest. Their level of pain mirroring was significantly reduced when observing painful stimuli being applied to individuals who had played unfairly. Thus, their level of pain activation was not automatic in the sense of being purely stimulus driven. Rather, it was modulated by internal preferences acquired from information about the targets. A similar result was obtained by Lamm et al. (2007), who found that subjects have a weaker empathetic response in pain-related areas when they know that the pain inflicted on another is useful as a cure.

However, the fact that mirroring can be modulated does not imply that it is a constructive activity comparable to creating an imagined scenario or adopting another person's perspective. Modulation of pain responses is inhibitory activity, something much less complex than the construction of an imagined scenario. It is the constructional aspect of many instances of empathizing I mean to highlight here. Mirroring is subject to modulation, but this doesn't make it a constructive or effortful activity, like some form of empathizing appears to be. In view of these features of the second type of empathizing, I shall call it *reconstructive* empathy (cf. Vignemont (2008)).

Another argument for this second route to empathy proceeds as follows. Assume that this kind of empathizing involves adopting the perspective of the empathic target. It is widely thought that such perspective-taking (arguably a form of simulation) is a crucial part of mindreading, or 'theory of mind' (ToM). We can then argue from functional neuroimaging data about theory of mind that this kind of empathizing is probably not the same as mirroring, because the brain regions subserving ToM have minimal overlap with either motoric mirror areas or areas involved in the mirroring of sensations or emotions. Of course, we have previously argued that mirroring is not subserved by a unique set of brain regions. In principle, then, areas involved in ToM might be mirror areas. This is possible in principle, but there is no evidence to support it. Thus, if empathizing is involved in these other types of mindreading—for example, attribution of beliefs and thoughts—it is likely to be a different type of empathizing process than mirroring.

Which brain regions are implicated in the mindreading of beliefs and other propositional attitudes? According to a number of researchers, they include the medial prefrontal cortex (MPFC), the temporo-parietal junction (right and left), and the temporal poles. Some authors contend that one area in particular, the right temporo-parietal junction (RTPJ), is specifically involved in tasks concerning belief attribution (Saxe & Kanwisher (2003); Saxe & Powell (2006); Saxe (2006)). Assuming that these brain regions are indeed involved in mindreading, what reason is there to suspect a connection between them and empathizing? What is the connection, after all, between mindreading and empathizing—especially 'reconstructive' empathizing?

According to the simulation approach to mindreading, especially as developed in my *Simulating Minds* (Goldman (2006a), there is a very tight connection. In what I call 'high-level' simulation (Goldman (2006a): ch. 7), mindreading another person's mental state involves an attempt to replicate or re-experience the target's state via a constructive process. Exploiting prior information about the target, the mindreader uses 'enactment imagination' to reproduce in his own mind what might have transpired, or may be transpiring, in the target. This coincides with the reconstructive type of empathizing proposed here. The only difference is that mindreading involves an additional final step in which one or more of the constructed mental states are categorized (commonly, in terms of both mental type and propositional content) and assigned to the target. This final stage—especially the categorization element—may be absent in empathizing.

## 3.4 A Possible Neural System Subserving Reconstructive Empathizing

Is there a neural system that subserves a process of reconstructive empathizing? Let us reconnoiter the subject by starting at what seems like a great distance: episodic memory. Episodic memory allows individuals to project themselves backwards in time and recollect aspects of their previous experience (Tulving, 1983; Addis et al., 2007). A growing number of investigators, however, have begun to approach episodic memory in a broader context, one that emphasizes people's ability both to re-experience episodes from the past and also imagine or 'pre-experience' episodes that may occur in the future (Atance & O'Neill (2005); D'Argembeau & Van der Linden (2004); Gilbert (2006); Klein & Loftus (2002); Schacter & Addis (2007); Schacter, Addis, & Buckner (2007); Buckner & Carroll (2007)). Evidence for a linkage between representations of past events and future events initially comes from studies of patients with episodic memory deficits. Tulving's (1985) patient K.C. suffered from total loss of episodic memory due to damage to the medial temporal and frontal lobes. K.C. was also unable to imagine specific events in his personal future, despite no loss in general imagery abilities. A second amnesic patient, D.B., also exhibited deficits in both retrieving past events and imagining future events (Klein & Loftus (2002)). D.B.'s deficit in imagining the future was also specific to his personal future; he could still imagine possible future events in the public domain (e.g. political events and issues). In general, projecting one's thoughts backward or forward in time is referred to as 'mental time travel.'

Hassabis et al. (2007) examined the ability of five amnesic patients with bilateral hippocampal damage to imagine novel experiences (see the summary by Schacter et al. (2007)). The imaginary constructions by four of the five patients were greatly reduced in richness and content compared with those of control subjects. Since this study did not specifically require patients to construct scenes pertaining to future events, they seem to suffer from a more general deficit to construct novel scenes. Recent neuroimaging studies provide insight into whether common brain systems are used while remembering the past and imagining the future. In a PET study by Okuda et al. (2003) participants talked freely about either the near or distant past or future. The scans showed evidence of shared activity during descriptions of past and future events in a set of regions that included the prefrontal cortex and parts of the medial temporal lobe—namely the hippocampus and the parahippocampal gyrus.

Drawing on these and related studies, Buckner and Carroll ((2007); see also Schacter et al. (2007); Schacter et al. (2008); Schacter & Addis (2009)) have proposed a core brain system that subserves as many as four forms of self-projection. These include remembering the past, thinking about the future (prospection), conceiving the viewpoint of others (theory of mind), and navigation. What these mental activities all share is a shift of perspective from the immediate environment to an alternative situation. All four forms rely on autobiographical information and are constructed as a 'perception' of an alternative perspective. (This brain system also goes under the label of 'the default network.')

The hypothesized core brain system involves frontal lobe systems traditionally associated with planning and medial temporal-parietal lobe systems associated with memory. How does theory of mind (ToM) fit into this picture neuroanatomically? Buckner and Carroll suggest that Saxe & Kanwisher's (2003) findings on the role of right TPJ in ToM provide further evidence that the core system extends to ToM. In the Saxe & Kanwisher (2003) study, individuals answered questions about stories that required participants to conceive a reality that was different from the current state of the world. In one condition the conceived state was a belief; in the other, it was an image held by an inanimate object (e.g. a camera). Conceiving of the beliefs of another person strongly activated the network shared by prospection and remembering, whereas the control condition did not. Buckner and Carroll also cite Gallagher & Frith's (2003) proposal that the frontopolar cortex contributes to ToM. In particular, the paracingulate cortex, the anterior-most portion of the frontal midline, is recruited in executive components of simulating others' perspectives. Thus, the Buckner-Carroll suggestion is that the core brain system is used by many diverse types of task that require mental simulation of alternative perspectives, and this includes thinking about the perspectives of other people.

Shanton (unpublished) follows up the hypothesis of Schacter, Addis, Buckner, and Carroll by identifying an assortment of experimentally confirmed parallels between episodic memory and ToM. She begins by explaining how each can be understood as a form of 'enactment imagination,' in the sense of Goldman ((2006a): chs. 2 and 7).

Enactment imagination is a species of imagination in which one tries to match a mental state or sequence of mental states in another by recreating or pre-creating this state or states in oneself. Shanton (unpublished; Shanton & Goldman 2010) argues that if the same type of simulation strategy is used for both episodic memory and mindreading tasks, there should be parallels in terms of various cognitive parameters. She reviews evidence of several such parallels, including (1) their developmental timeline and (2) their susceptibility to egocentric biases.

Consider first the fact that episodic memory and mindreading share a developmental timeline. According to Tulving (2001), episodic memory retrieval emerges around the age of 4 years. This is confirmed by Perner & Ruffman (1995), who had children between 3 and 6 years of age complete both free recall and cued recall memory tasks. These tasks tap different types of memory abilities. In cued recall tasks, semantic information is quite rich, whereas in free recall tasks, where no explicit retrieval cues are given, such information is relatively poor. Free recall tasks cannot be successfully answered without episodic memory. Perner and Ruffman found that only 4–6-year-old children, not 3-year-olds, could succeed on free recall tasks, supporting the hypothesis that episodic memory retrieval emerges around age 4. This corresponds to the traditional timeline for success in advanced mindreading tasks, such as (verbal) false-belief tasks.

Next consider the susceptibility of both high-level mindreading and episodic memory retrieval to egocentric biases. One example of egocentric bias is the 'curse of knowledge' (Camerer et al. (1989)). This is the tendency to proceed as if other people know what you do, even when you have information to the contrary. In the Camerer et al. study, well-informed people were required to predict corporate earnings forecasts that would be made by other, less-informed people. The better-informed people stood to gain if they disregarded their own knowledge when making predictions about the less-informed people, who they *knew* lacked the same knowledge. Nonetheless, they failed to disregard their own knowledge completely, letting it 'seep' into their predictions. Simulationists would say that the predictors, while attempting to imagine themselves in the shoes of the predictees, allowed their own knowledge to 'penetrate' their imaginative construction. In other words, their own genuine mental states were not excluded, or quarantined, from the construction, despite the fact that good (i.e. accurate) simulation requires such quarantining. Quarantine failure is extremely common in (high-level) mindreading. For example, Van Boven & Loewenstein (2003) asked participants to predict states like hunger and thirst in a group of hypothetical hikers lost in the woods with neither food nor water. Their predictions were solicited either before or after they vigorously exercised at a gymnasium. In the case of post-exercise participants, the combined feelings of thirst and warmth were positively associated with their predictions of the hikers' feelings. Here too there is apparent failure to quarantine one's own concurrent states while mindreading hypothetical targets.

Quarantine failure is found in episodic memory. A vivid illustration is from Levine's (1997) study of subjects' memories for their own past emotion states. During the 1992 presidential race, Levine first asked a group of Ross Perot supporters about their emotions immediately after Perot withdrew from the race in July, and later asked them again in November, after they had switched their allegiances to other candidates. Although in July they described themselves as very sad, angry, and hopeless, by November they remembered experiencing much lower levels of emotion (in July). Apparently their November memories were being influenced by their current attitudes toward Perot. Their episodic memories were constructions that were partly influenced, or colored, by the way they felt at the time of memory 'retrieval'.

Shanton argues that the best explanation of these similarities is that the two processes implement the same cognitive strategy, and she argues (based on additional evidence) that this strategy is enactment imagination. This is a different form of simulation than mirroring.

How does enactment imagination differ from mirroring? In the case of mirroring processes, the default upshot is the successful production of a match between the sender state and the receiver state. Disgust in a sender is reproduced, with reasonable accuracy, by disgust in the observer. In the case of enactment imagination, by contrast, the prospects for successful correspondence are much more tenuous. They heavily depend on the vicissitudes of prior information, construction and/or elaboration. In the case of mindreading, the vicissitudes of prior information are particularly important. If one doesn't have accurate and relevantly complete information about the prior mental states of the target, attempts to put oneself in that person's mental shoes in order to extrapolate some further mental state have relatively shaky prospects for success.

In my view, it isn't entirely clear that the same core brain system described by Buckner, Carroll, Schacter, and colleagues includes ToM, or mentalizing. For example, in describing their core system, Buckner and Carroll say that it extends to lateral parietal regions located within the inferior parietal lobule 'near' the temporo-parietal junction. But being *near* TPJ may not be sufficient to identify this area as a locus of mentalizing activity. However, my brief for a simulation system that leads to both mindreading and empathizing via *reconstruction* rather than mirroring does not depend essentially on neuroanatomical evidence. If the specific core brain system hypothesized by Buckner, Carroll, and Schacter does not extend to mindreading and empathizing, this would still be compatible with there being a constructive, or reconstructive, species of empathizing. If the core brain system does subserve mindreading and empathizing, that is just gravy.

## 3.5 Output Profiles of the Two Routes to Empathy

The topic of the last three sections has not been states of empathy *per se* but different *routes* to empathy. Routes to empathy are species of mental activity that (often) lead to empathic states, where empathic states are defined as indicated in section 3.1 (with possible modifications considered there). The next natural question to ask is how successful or unsuccessful are the two different routes in generating empathic states, that is, states that exemplify substantial isomorphism to those of their targets. How do they fare in comparative terms? Are there characteristic differences in the empathic outputs of the two different routes?

The question of comparative success or accuracy can be decomposed into several sub-questions. First, one can ask about the *reliability* of a route or method. Of the states produced by a given method, how many are genuinely isomorphic to those of the target?[2] Second, one can ask about the *fecundity* of a route or method. For each application of a method, how many isomorphic states (on average) are produced in the empathizer? It should not be assumed that each application generates precisely one output state. Either of the two methods may generate more states (per use) than the other, and such greater fecundity may be important because it is associated with greater intersubjective understanding.

Restricting ourselves initially to the reliability question, the issue resolves into further sub-questions, because each state has more than one dimension and we can ask with respect to each dimension whether a given method produces output states that resemble the target on that dimension. Vignemont (2010) distinguishes four main dimensions of emotional states: (1) the *type* of state, (2) the *focus* (object) of a state, (3) the *functional role* of a state, and (4) the *phenomenology* of a state. A given route or method of empathizing might be more reliable than another route with respect to some of these dimensions but less reliable with respect to others.

For reasons previously sketched, it seems likely that the mirroring method of empathizing is more reliable than the reconstructive method when it comes to the *type* of emotional state. Mirroring, by its very nature, is a highly reliable method of state generation, one that preserves at least the sameness of mental-state type (e.g. pain, disgust, fear). There is no comparable guarantee (or near-guarantee) in the case of the reconstructive method. Outputs of a constructive or reconstructive method depend heavily on the pretend inputs that the empathizer uses, and the accuracy of these inputs can vary widely depending on the quality of her background information. In short, in terms of reliability with respect to *type*, the mirroring route seems superior to the reconstructive route.

What about the focus dimension of the state: what the emotion or other state is *about*? As Vignemont argues, the mirroring method does not seem to be so helpful in this regard, whereas the reconstructive method is (or might be). Vignemont's example is seeing a smiling stranger on a train. Seeing the smile prompts a happy state in the observer by the mirroring route. But the object of the stranger's happiness remains

[2] The term 'reliability' is not used here in exactly the same sense in which it is used in epistemology, because we are not discussing the formation of true or false beliefs. Instead, we are discussing how much, or to what degree, one mental state is isomorphic to (resembles) another. Degrees of reliability are to be computed in terms of proportions of isomorphic versus non-isomorphic features (or something along these lines).

undisclosed by mirroring. Mirroring reproduces in the observer only happiness, not happiness *about* X or *about* Y. On the other hand, argues Vignemont, reconstructive empathizing can be helpful with respect to focus. By adopting the target's perspective, an empathizer can figure out what the object's emotion is about, or directed at, at least when appropriate information is available. Thus, reconstructive empathizing seems to be superior to mirroring in this regard.

Vignemont includes an *intensity* dimension for output states, which she subsumes under phenomenology. Although I agree that an intensity dimension is relevant here, it is not clear that it should be confined to the sphere of phenomenology. As previously indicated, mirroring states often fail to reach the threshold of consciousness, so they may have no phenomenology at all. Does this mean that they have zero intensity? This would be an unsatisfactory inference because unconscious states certainly have important functional properties, including tendencies to influence behavior. On the other hand, what alternative measure of intensity should be selected? Should some measure of neural activity be used? Which one? In any case, once a measure of intensity is chosen, the question is whether the mirroring method or the reconstructive method is more reliable, that is, which tends to produce mirrored states with greater isomorphism? It isn't obvious (to me) what the answer is; this question invites more research.

Vignemont regards the reconstructive method as superior to the mirroring method, but since she doesn't herself draw the reliability/fecundity distinction, it is an open question whether the intended superiority is supposed to hold for both reliability and fecundity, or for fecundity only. She writes:

> Low-level [mirroring] empathy does not meet the condition of isomorphism [because it is limited to the *type* of emotion, and does not go beyond that]. Emotional sharing may be more exhaustive in high-level empathy. (2008)

To support this idea she considers a case of a woman learning that a friend is pregnant. Since the empathizer knows how much the friend wanted a child, she puts herself in the friend's shoes and realizes how happy she must be. She feels happy with her. The inputs in such a case of reconstructive empathizing are more complex than the inputs to mirroring empathizing, and this allows one to fill out the target's mental states more fully, or in greater detail. Continuing with the pregnancy example, the empathizer pretends that she is pregnant and that she wants a child, which leads her to feel happy. Her emotional state is *about* the pregnancy; it has the same focus as the friend's emotional state. The mirroring method, says Vignemont, 'isolates' a mirrored emotion from the rest of the target's mental life. It does not provide a fine-grained sharing of states based on a common network of associated mental states. Reconstructive empathizing does provide this.

Suppose Vignemont means to say that the reconstructive method is superior in both reliability and fecundity. I would be prepared to concede the fecundity part, because the reconstructive method is obviously capable of generating more and more detailed isomorphic states than mirroring. But is it more reliable? I am skeptical. Vignemont ignores two types of error to which only the reconstructive method is liable. The first type of error is an error of omission: omitting relevant inputs because of ignorance. If an (attempted) empathizer in the pregnancy case is unaware that her friend is pregnant or is unaware that she wants a child, application of the reconstructive method is unlikely to produce *correct* details involving the target emotion. The second type of error is an error of commission. As reviewed above, there is substantial evidence that when people try to simulate the mental states of others, they often fail to 'quarantine' their own genuine states, allowing such states to seep into the simulation process when they don't properly belong there (because the target isn't in them). This results in 'egocentric biases' in the simulation process. Both types of errors can substantially reduce the reliability of the reconstructive method, so I cannot concur with Vignemont's rosy appraisal of it. An assessment of the comparative reliability of the two methods needs more work. Nonetheless, it is good to have this problem placed squarely on the table; it deserves attention.

I have argued that there are two distinct routes to empathy, the mirroring route and the reconstructive route. It is possible, however, that the reconstructive route also involves mirror neurons. This is suggested by Iacoboni and colleagues (Iacoboni, this volume; Uddin et al. (2007)). Iacoboni (this volume) reports the recent discovery of mirror neurons in several new areas, including the amygdala, hippocampus, parahippocampal gyrus, and entorhinal cortex. He suggests that these mirror neurons may underpin what I earlier called 'high-level' mindreading and empathy (Goldman, 2006a), which correspond to what is here called reconstructive empathy.[3] Thus it is possible that even reconstructive empathy is mediated by neurons with mirror properties. Note, however, that this would not necessarily undercut the distinction between mirroring and reconstructive processes. As standardly conceived, mirror processes are automatic processes generated by observation. In addition, neurons with mirror properties might also participate in such an effortful process as imagination (see Uddin et al. (2007), box 3), a key component of reconstructive empathy. These ideas require further investigation.

[3] Notice that some of the areas containing mirror neurons mentioned by Iacoboni are the same as midline areas mentioned by Okuda et al. (2003) in their study of constructive imagination, specifically, the hippocampus and the parahippocampal gyrus.

# 4

# Within Each Other:

## Neural Mechanisms for Empathy in the Primate Brain

*Marco Iacoboni*

## 4.1 Introduction

Empathy is commonly defined as the ability to understand and share the feelings of another. It is obviously a very complex ability. What are the neurophysiological mechanisms that underlie empathy? For years, nobody dared to investigate this issue. The main reasons were two. First, the study of the brain mechanisms associated with emotion and emotional understanding is relatively recent. Until approximately 20 years ago, the study of the neural systems associated with higher functions was focused exclusively on 'cold' cognitive processes. The dominant metaphor was 'the mind as a computer.' The study of emotions—especially complex social emotions—clearly did not fit in the prevalent paradigm. Second, even after emotions became a popular topic in cognitive neuroscience, mostly thanks to the influential work of Antonio Damasio, the neural mechanisms of empathy remained largely unexplored. This was likely due to the perceived complexity of empathy. Indeed, the complexity of a phenomenon is generally considered an obstacle for the study of its neural correlates, especially in single cell recordings. While neurophysiologists are able to study brain activity at its most exquisite spatial and temporal resolution, that is, the spiking activity of single cells, they also tend to study this activity in relation with relatively simple phenomena, such as the perception of individual sensory stimuli or the planning and execution of relatively simple actions. For this reason, neurophysiological data on empathy were virtually nonexistent until a few years ago. In recent years, however, a new wave of studies has investigated the links between empathic behavior and brain activity. The recent studies have been inspired by the discovery of mirror neurons in the macaque brain. These cells, which I describe in detail in the next section of the chapter, have *physiological properties* that are ideal to facilitate empathy. Indeed, the properties of mirror neurons seem to map extremely well onto emotional contagion, a phenomenon studied for decades by psychologists (Hatfield et al. (1994)). Most