Figure 1 (Eslinger et al.).     Results of fMRI study of normal volunteers making explicit moral judgments, showing activation primarily of frontopolar cortex, medial frontal gyrus and related regions

social bonds have been traditionally conceived to rely on two main streams of processing (Eslinger 1998): one more cognitively oriented and dependent on the dorsolateral prefrontal cortex, and the other more emotion-oriented and mediated by anterior temporal, limbic, and orbitomedial frontal cortical systems. In a neuropsychological study of humans with acquired lesions of the frontal lobe (Grattan et al. 1994), comparisons between quantitative measures of cognitive flexibility and empathy indicated that left and right dorsolateral frontal lesions caused deficits in both cognitive flexibility and empathy (with correlations ranging as high as r = .81). In contrast, orbitofrontal lesions profoundly affected empathy but did not alter cognitive flexibility, suggesting different mechanisms of pathophysiology for empathic change that may have its basis in deficient automatic tagging of salient stimuli or evocation of autonomic-visceral states. The latter pattern has also been reported in patients with lesions of the insula and deep white matter of the frontal lobe, likely disconnecting orbitofrontal, amygdala, and insula interconnections. Despite the lack of automatic emotional responsiveness, these individuals can be trained to use cognitive mechanisms for recognition of salient facial, vocal, and body expressions, and for the verbal expression of their concern. Thus, frontal regulation of social conduct favors the view that the prefrontal cortex may need further functional fractionation, with polar, orbital, medial, dorsolateral, and deep white matter sectors mediating distinct, but complementary roles in the emotional-cognitive regulation of social behavior. In our experience, early damage to polar, mesial, and dorsolateral regions disrupts the developmental acquisition of automatic and learned empathy, moral judgment, and interpersonal conduct (e.g., Eslinger et al. 1992; 1997).

The cerebral correlates of specific dimensions of moral and empathic behaviors need further study (Grattan & Eslinger 1989). P&deW's article offers a solid conceptual frame against which specific hypotheses and models can be put to empirical testing in normal individuals, as well as in patients with a variety of neuropsychiatric disorders which express themselves as disorders of empathy, social, and moral behavior.

# The mirror matching system: A shared manifold for intersubjectivity

Vittorio Gallese, Pier Francesco Ferrari, and Maria Alessandra Umiltà

*Istituto di Fisiologia Umana, Università di Parma, I-43100 Parma, Italy.*
**vittorio.gallese@unipr.it     ferrari@biol.unipr.it     maumilta@unipr.it**

**Abstract:** Empathy is the phenomenal experience of mirroring ourselves into others. It can be explained in terms of simulations of actions, sensations, and emotions which constitute a shared manifold for intersubjectivity. Simulation, in turn, can be sustained at the subpersonal level by a series of neural mirror matching systems.

The article by Preston & de Waal (P&deW) has the merit of raising the issue on how different and multifarious social behaviors such as emotional contagion, helping behaviors, imitation, and cognitive empathy may be reconciled within a unified explanatory framework. Their proposal is that empathy can constitute such a framework, provided that it is "construed broadly to include all processes that rely on the perception-action mechanism" (see target article, sect. 1.1.3, Fig. 1 caption).

Unfortunately, the discussion of these interesting issues is, in our view, most of the time too vague, and sometimes even confounding, so that it is difficult to draw any firm conclusion on the validity of the proposed model.

The definition of empathy given by P&deW seems to oscillate between a "superordinate category" of behaviors, and a process underpinning the same different types of behavior. They introduce the Perception-Action Model (PAM), which is presented as a "superordinate class," capable of including and subsuming different types of behaviors and effects, among which empathy is also listed (see their Fig. 1). This makes it almost impossible to understand whether in the authors' view empathy and PAM are the same thing or not. Furthermore, the term PAM is used interchangeably to denote either a model (as the acronym seems to suggest) or a basic (neurophysiological?) level of description. Such a use of the term makes several statements hard to understand. How can a model possibly *interact* with specific brain functions? (e.g., see the Abstract where it is stated that: "The interaction between the PAM [Perception Action Model] and prefrontal functioning can also explain different levels of empathy.") Models should *explain*, or *interpret* functions, not interact with them.

The broad and general sense attributed by P&deW to the term empathy fails to provide a coherent picture of the mechanism that is at the basis of their model. There is confusion here – if not even a category mistake – about the chosen *level of explanation.*

In our opinion, the term "empathy" makes sense only if used to denote a phenomenological level of description: the one responsible for the sense of similarity that we experience anytime we confront ourselves with other human beings, and sometimes even with animals. Empathy is deeply grounded in the experience of our lived-body, and it is this experience that enables us to directly recognize others, not as bodies endowed with a mind, but as *persons* like us. Actions, sensations, and emotions experienced by others become meaningful to us because we can *share* them with others. How can such a sharing of experiences be possible? We need to introduce a second level of description, one pertaining to the functional mechanism that enables such an experience to occur. This functional mechanism may be constituted by *simulation.* Simulation mechanisms are a possible way for a given organism to control and model its performances. By modeling a given process, our brain provides a *simulated representation* of the same process that can be used to produce it, on the one hand, and to decode it when performed by someone else, on the other. We suggest that these "as if" simulation mechanisms may subsume a wide range of processes as diverse as action perception and imitation (as simulation of the observed action; see Gallese 2000a; 2000b; 2001; see also Rizzolatti et al. 2001), emotion perception (as simulation of the perceived emotion; see Adolphs 1999; Adolphs et al. 2000;

Gallese 2001), and mindreading. Simulation theory in fact holds that we understand others' thoughts by *pretending* to be in their "mental shoes," and by using our own mind/body as a model for the minds of others (Gallese & Goldman 1998; Goldman 1989; Gordon 1986; Harris 1989).

Is there a further level of description that can provide a common and coherent explanatory frame for all these different simulation mechanisms? We propose, yes: such a level could be represented by the neural matching system constituted by mirror neurons (Gallese et al. 1996; 2002; Rizzolatti et al. 1996; Umiltà et al. 2001; see also Rizzolatti et al. 2001) – or by equivalent neural systems described in the human brain (Fadiga et al. 1995; Iacoboni et al. 1999; Nishitani & Hari 2000). Mirror neurons could underpin a direct, automatic, nonpredicative, and noninferential simulation mechanism, by means of which the observer would be able to recognize, understand, and imitate the behavior of others. The authors maintain that "mirror neurons . . . provide concrete cellular evidence for the shared representations of perception and action" (see target article, sect. 3.1). They fail, nevertheless, to draw the correct conclusions from such a statement. It is true, as they argue, that mirror neurons do not produce per se any empathy. However, if an action-perception matching is crucial for the production of empathy, as the authors suggest, mirror neurons represent the most parsimonious neural system so far described, enabling such a matching to occur. The trick here is not to confound the phenomenal aspect of behavior, its functional level of description, and the neural mechanism at its base.

Preliminary results suggest that a mirror matching system could be at the basis of our capacity to perceive in a meaningful way, not only the actions, but also the sensations and the emotions of others (see Gallese 2001). Single neuron recording experiments in humans have demonstrated that the same neurons become active when the subject either feels pain or observes others feeling pain (Hutchison et al. 1999). Furthermore, a recent fMRI study has shown that the amygdala becomes active not only during the observation, but also during the *active expression* of facial emotions, especially when imitation is involved (Carr et al. 2001).

In conclusion, these recent findings suggest that a neural matching system is present also in a variety of *apparently* non-motor-related human brain structures. Thus, different *simulation mechanisms* are applied in different domains, being sustained by a mirror-matching, dual-mode of operation (*action-driven* and *perception-driven*) of given brain structures. We propose that such simulation mechanisms may constitute altogether a *shared manifold* of intersubjectivity (see Gallese 2001).

# Cognitive empathy presupposes self-awareness: Evidence from phylogeny, ontogeny, neuropsychology, and mental illness

Gordon G. Gallup, Jr. and Steven M. Platek

*Department of Psychology, State University of New York at Albany, Albany, NY 12222.* **gallup@csc.albany.edu      steven_platek@yahoo.com**
**http://www.albany.edu/~gallup      http://www.albany.edu/~sp5681**

**Abstract:** We argue that cognitive empathy and other instances of mental state attribution are a byproduct of self-awareness. Evidence is brought to bear on this proposition from comparative psychology, early child development, neuropsychology, and abnormal behavior.

The subcategory of empathy that Preston & de Waal (P&deW) identify as cognitive empathy represents an instance of a more general phenomenon known as mental state attribution. Twenty years ago a model was formulated that stipulates self-awareness as a necessary condition for making inferences about mental states in others (Gallup 1982). According to the model, inferential knowledge of mental states in others builds on a knowledge of mental states in oneself. Organisms that can become the object of their own attention (i.e., recognize themselves in mirrors) can use their experience to model/infer comparable experiences in others. Because humans share similar receptor mechanisms and brains that are organized in roughly the same way, there is bound to be considerable overlap between their experiences. Moreover, people that have access to their own mental states and take note of their relationship to various external events, have a means of making inferences about mental states in others. Knowledge of self, in other words, paves the way for achieving an inferential knowledge of cognitive states in others.

Since the ability to recognize oneself in a mirror varies as a function of species, age, neurological status, and mental illness (Gallup et al. 2002), this model can be tested in each of these domains. According to the model, variation in mirror self-recognition ought to predict comparable variation in the ability to take into account how others feel, and variation in being able to accurately infer what they want, know, or intend to do. In instances in which self-recognition is deficient or absent, there should be a corresponding deficiency or absence of mental state attribution.

Thus, whereas chimpanzees that recognize themselves in mirrors show evidence of cognitive empathy (Povinelli et al. 1992a), rhesus monkeys, which fail tests of self-recognition, show no evidence of cognitive empathy (Povinelli et al. 1992b). Monkeys seem incapable of taking into account what other monkeys may or may not know, want, or intend to do. Indeed, monkeys may not even know what they know (Cheney & Seyfarth 1990).

The same holds for humans. People who are incapable of recognizing themselves in mirrors are typically unable to make inferences about mental states in other people. For instance, it is only after children learn to recognize themselves in mirrors (usually between 18 and 24 months of age) that they begin to show evidence of being able to take into account what other people are seeing or feeling (e.g., Carruthers & Smith 1996). Likewise, the emergence of prosocial and altruistic behaviors in children is related to the age at which they show self-recognition (Johnson 1982).

There is a growing evidence that self-recognition is localized in the brain. Patients who are incapable of identifying their own faces often have damage to the right prefrontal cortex (e.g., Spangenberg et al. 1998). Breen (1999) described a patient with damage restricted to the right prefrontal cortex that could recognize other people in a mirror, but insisted that his own reflection was someone else. Keenan and Wheeler (in press) summarize a number of neuropsychological studies which show both right hemispheric lateralization and localization of self-recognition in the prefrontal cortex. As support for the proposition that mirror self-recognition is a valid index of self-awareness, it is important to note that self-evaluation and autobiographical memories are also localized in the right prefrontal cortex (see Keenan et al. 2000).

Consistent with the model, the same part of the brain that is important for self-recognition also appears to be crucial for mental state attribution. In contrast to patients with brain damage elsewhere, Stuss et al. (2001) found that those with lesions restricted to the right frontal lobes showed deficits in visual perspective taking and detecting deception. Likewise, Happe et al. (1999) report that patients with damage to the right hemisphere evidence an impaired ability to interpret mental state attribution narratives and fail to appreciate instances of humor that require making mental state attributions.

Deficits in mirror self-recognition are also characteristic of a number of psychological disorders (Gallup et al., in press). Schizophrenics often react to themselves in mirrors as though they were seeing other people (Harrington et al. 1989). Similarly, people who score high on measures of schizotypal personality show impairments in self-face recognition (Platek & Gallup 2002). There is mounting evidence that schizophrenia is associated with frontal