Routledge
Taylor & Francis Group

# Motor ontology: the representational reality of goals, actions and selves

## Vittorio Gallese & Thomas Metzinger

ABSTRACT   *The representational dynamics of the brain is a subsymbolic process, and it has to be conceived as an "agent-free" type of dynamical self-organization. However, in generating a coherent internal world-model, the brain decomposes target space in a certain way. In doing so, it defines an "ontology": to have an ontology is to interpret a world. In this paper we argue that the brain, viewed as a representational system aimed at interpreting the world, possesses an ontology too. It decomposes target space in a way that exhibits certain invariances, which in turn are functionally significant. A challenge for empirical research is to determine which are the functional regularities guiding this decomposition process. What are the explicit and implicit assumptions about the structure of reality, which at the same time shape the causal profile of the brain's motor output and the representational deep structure of the conscious mind arising from it (its "phenomenal output")? How do they constrain high-level phenomena like conscious experience, the emergence of a first-person perspective, or social cognition? By reviewing a series of neuroscientific results, we focus on the contribution the motor system makes to this process. As it turns out, the motor system constructs goals, actions, and intending selves as basic constituents of the world it interprets. It does so by assigning a single, unified causal role to them. Empirical evidence now clearly shows how the brain actually codes movements and action goals in terms of multimodal representations of organism–object relations. Under a representationalist analysis, this process can be interpreted as an internal representation of the intentionality relation itself. We try to show how such a more complex form of representational content, once it is in place, can later function as the building block for social cognition and a for more complex, consciously experienced representation of the first-person perspective as well. The motor system may therefore play a decisive role in understanding how the functional ontology of the human brain could be gradually extended into the subjective and social domains.*

## 1. What is a motor ontology?

One useful way to look at the brain is to describe it as a dynamic representational system. Every representational system presupposes an *ontology*: a set of assumptions about what the elementary building blocks of the reality to be represented actually are. By necessity, it constructs *primitives*. For example, many natural languages,

*Vittorio Gallese, Dipartimento di Neuroscienze, Sezione di Fisiologia Umana, Università di Parma, Via Volturno 39, I-43100 Parma, Italy, email: vittorio.gallese@unipr.it*
*Thomas Metzinger, Philosophisches seminar, Johannes Gutenberg-Universität, Mainz, Germany, email: metzinger@uni-mainz.de*

viewed as representational systems, typically assume that extralinguistic reality is constituted by objects, properties, and relations. Their underlying ontology then frequently carries over into folk-psychological contexts, for instance, by influencing the way in which we naively describe our own mental states in everyday life situations. A quantum mechanical description of reality in physics, on the other hand, may successfully operate under a completely different set of metaphysical assumptions. It represents reality under a *different* ontology, because it uses different primitives. An interesting but frequently overlooked fact is that the human brain possesses an ontology too, because it makes assumptions about what the relevant and what the irreducible building blocks of external reality are. By metaphorically speaking of the brain as "making assumptions" (or as "arriving at conclusions," etc.; see below), of course, we do not mean to imply that the brain is a cognitive or epistemic agent, or even that it internally uses propositional formats and symbolic coding principles. *Persons* make assumptions, and brains are only *subpersonal* functional modules of whole persons. The representational dynamics of brains is a sub-symbolic process, defying the distinction between syntax and semantics, and it has to be conceived as an "agent-free" type of subpersonal *self-organization* [1]. However, in generating a coherent internal world-model, brains *decompose target space* in a certain way. The functional regularities guiding this decomposition process are by no means arbitrary. We are interested in the *invariances* guiding it, and this is what we mean by investigating the brain's "ontological assumptions." However, we will here take no position at all on the ontological commitments of cognitive neuroscience or the metaphysics generated by scientific theories in general—in principle, it could turn out that brains are not physical objects in the traditional materialist sense, but thoughts in the mind of God. This would not touch the point we are trying to make. Even brains of this kind would exhibit exactly the same internal structure within their reality-model, although the substrate and reference of this model would then be entirely different from what the majority of scientists assume today.

In this paper we ask: what is the *functional ontology* of the human brain? What are the explicit and implicit assumptions about the structure of reality, which at the same time shape the causal profile of its motor output and the representational deep structure of the conscious *mind* arising from it (its "phenomenal output"), and how do they constrain high-level phenomena like phenomenal experience or social cognition?

The target of this paper will be to analyze the contribution, which the *motor system* makes to some fundamental aspects of the brain's ontology. In particular, we want to focus on the issue of how a dynamical, complex and self-organizing physical system could arrive at the conclusion that goals, actions, and intending selves actually belong to the basic constituents of the world that it is internally modeling [2].

Second, although we are interested in how this process constrains and enables high-level phenomena like subjectivity and phenomenal experience, we do not want to imply a solution to the problem of consciousness as such. As will become obvious in the course of this paper, the representational dynamics we describe cuts *across* the

border between phenomenal and non-phenomenal states in an interesting way. We will start with a first set of data that has led us to the reconceptualization of issues and will present the conclusions later.

## 2. Empirical evidence

During the last decades neuroscientific research has started to confront itself with the intentional dimension of behavior. The study of motor functions, carried out by investigating different levels of its hierarchical organization, when tackling the functions of premotor areas, has quite unexpectedly led to the discovery of "goal-related" neurons. Soon we will be able to be more specific: "goal-related" neurons are precisely all those neurons contributing to goal *representations* (to be defined in Section 3), e.g. to representations of possible, reward-producing actions as seen from no particular perspective.

By electrophysiological means the area called F5 (Matelli *et al.*, 1985), which occupies the rostral most part of the ventral premotor cortex of the monkey, has been delineated to contain a distal hand and mouth movement representation (Gentilucci *et al.*, 1988; Hepp-Reymond *et al.*, 1994; Kurata & Tanji, 1986; Rizzolatti *et al.*, 1981, 1988). The most interesting aspect of F5 neurons is that they code movement in quite abstract terms. What is coded is not simply a parameter such as force or movement direction, but rather the relationship, in motor terms, between the agent and the object of the action. F5 neurons become active only if a particular type of action (e.g. grasping, holding, etc.) is executed to achieve a particular type of goal (e.g. to take possession of a piece of food, to throw away an object, etc.). Particularly interesting in this respect are grasping neurons that fire any time the monkey grasps an object, regardless of the effector employed, be it the hand or the mouth (Rizzolatti *et al.*, 1988, 2000).

The metaphor of a "motor vocabulary" has been introduced (Rizzolatti *et al.*, 1988) in order to conceptualize the function of these neurons. This vocabulary collects various "words," each constituted by groups of neurons related to different motor acts. The hierarchical value of these "words" can be different: some of them indicate the general goal of the action (e.g. grasping, holding, tearing). Some other "words" concern the way in which a particular action has to be executed, e.g. to grasp with the index finger and the thumb (precision grip). Another group of "words" deals with the temporal phases in which the action to be performed can be segmented (e.g. hand aperture phase).

The presence in the motor system of a "vocabulary" of motor acts allows for a much simpler selection of a particular action within a given context. Either when the action is self-generated or externally generated, only a few "words" need to be selected. Let us imagine that a monkey is presented with a small object, say a raisin, within its reaching distance. If the motivational value of the stimulus is powerful enough to trigger an appetitive behavior, it will evoke a command for a grasping action; then the command will address a specific finger configuration, suitable to grasp the raisin in that particular situation. Within the context of a motor "vocabulary," motor acts can be conceived as a simple assembly of words, instead

of being described in the less economical terms of the control of individual move-ments (see Rizzolatti *et al.*, 1988). This leads to an interesting question: what level of abstraction governs the brain's functional ontology? What constrains the *level of granularity* of, for example, its motor ontology?

Since most grasping acts are executed under visual guidance, a relationship between the features of 3D visual objects and the specific "words" of the motor vocabulary has to be established. In this logic the appearance of a graspable object in the visual space will retrieve immediately the appropriate ensemble of motor "words." This process, in neurophysiological terms, implies that the same neuron must be able not only to code motor acts, but also to respond to the visual features triggering them. The first conclusion is that such neurons contribute to a multi-modal representation of an *organism–object relation*.

Indeed, a considerable percentage of F5 grasping neurons, "canonical neu-rons," respond to the visual presentation of objects of different size and shape in absence of any detectable movement (Jeannerod *et al.*, 1995; Murata *et al.*, 1997; Rizzolatti *et al.*, 1988, 2000). Very often a strict congruence has been observed between the type of grip coded by a given neuron and the size or the shape of the object effective in triggering its visual response. The most interesting aspect, how-ever, is the fact that in a considerable percentage of neurons the congruence is observed between the selectivity for a given type of executed grip and the selectivity for the visual presentation of objects that, although differing in shape, nevertheless all "afford" the same type of grip, which is identical to the motorically coded one.

The function of F5 canonical grasping neurons can hardly be defined in purely sensory or motor terms alone. At this stage object representations seem to be processed in *relational terms* (Gallese, 2000a,b). And this is another important conclusion about the causal microstructure of the brain's functional ontology: within the logic of such a neural network, a series of physical entities, 3D objects, are identified and differentiated not in relation to their mere physical appearance, but in relation to the effect of the interaction with an acting agent. This fact is a good first example for an important aspect of the brain's ontology, which has been frequently overlooked in the past. Intentionality—the directedness of mental states toward an object (Brentano, 1874/1973, p. 124)—is something that may or may not exist in the objective world-order. What we want to draw attention to is that there is good evidence that at least *the brain's own functional ontology* assumes that something like intentional acts in the classical sense of Brentano and Husserl actually do exist, because there are certain kinds of information which are specifically and exclusively coded under a distinct type of *internal* neural representation involving dynamic subject–object relations.

We think that these results should lead to a deep rethinking of what lies at the roots of intentional behavior. To speak of a neural level of representation that *has* to be qualified as "goal-related," however, doesn't necessarily imply that this level *per se* is also able to specify the action goal in a way that is *detached* from the means to achieve it. In the monkey brain microcosm so far explored, the goal of grasping an object is still almost completely overlapping with the action-control strategies. Action control actually equates to the definition of the action goal: the goal is

represented as a goal-state, namely, as a successfully terminated action pattern. However, the presence in the monkey brain of neurons coding the goal of grasping, regardless of which effector (hand or mouth) will eventually achieve it, in our opinion provides evolutionary evidence of the dawning of a more complex—and abstract—coding of intention and volition.

Even more so, if we consider the multimodal nature of canonical neuron responses, it becomes clear that—already in the monkey brain—the way the brain represents goal-directedness for the organism as a whole becomes increasingly abstract, and this is an advantage. It is no coincidence that such an intimate relation between object representation and action representation still holds in the human brain. Several brain imaging experiments have indeed shown that observation, silent naming, and imaging the use of man-made objects (but not that of animals) leads to the activation of the ventral premotor cortex (Chao & Martin, 2000; Grafton *et al.*, 1996; Martin *et al.*, 1996; Perani *et al.*, 1995), thus reflecting a process similar to that described above in the monkey.

In humans, in fact, we see the possibility to identify, specify, and think of goal-states that are somehow very distal to the chain of events leading to their fulfillment. For example, I can think of my future goal-state of having dinner at the restaurant without simultaneously bothering to specify all the intermediate steps leading to the achievement of the distal goal-state itself (calling the restaurant to reserve a table, driving to the restaurant, and the like). When I specify the goal-event of going to dine I focus on one event that is completely detached from the practical intermediate strategies required to make this event happen. All these strategies certainly clarify, contribute to delimit, and circumscribe the space of distal goal representations; nevertheless they do not need to be made explicit in order to individuate the specific point in this space, which unequivocally represents the relevant goal-state. In other words, the satisfaction conditions shaping the internal landscape of human goal-state space can be highly abstract, involving no information about the fine-grained causal microstructure necessary for their realization whatsoever.

The distal goal-state to dine at the restaurant, whose ultimate consequence will be to introduce food in my mouth (within a crowded place), can be decomposed into a chain of sub-events, each of which (phoning, driving the car, etc.) can hardly be seen as specific of the distal goal-state. The distal goal-state therefore seems to be attained through a series of multipurpose (I can phone my partner, I can drive to go to see a movie, I can eat at home) action schemata, which nevertheless have to be chained within a highly specific temporal sequence. Such a possibility to diversify the procedural sub-components while simultaneously filing them within the same distal goal-state representation could be interpreted as the result of an incredibly higher integrative capacity of the human brain.

## 3. Goals

We think that the data so far presented make it necessary to conceptually restructure the questions that have to be asked. Let us begin by defining some basic

notions. What is a goal? From a strict scientific point of view, no such things as goals exist in the objective world [3]. All that exists are goal *representations*, for instance, those activated by biological nervous systems. A goal representation is, first, formed by the representation of a certain state of the organism, of the world, or by the holding of a certain relation between the organism and a part of the world, e.g. another organism. Goal representations are representations of goal-*states*. Second, what makes such a state a goal-state is the fact that its internal representation is structured along an axis of valence: it possesses a value *for* the system.

Biological systems are systems under evolutionary pressure, systems having to predict future events that possess a high survival value. Therefore, the prehistory of representational goal-states is likely to be found in the reward system. Why is this so? Reward is the payoff of the self-organizing principles that functionally govern and internally model the organization of an open system such as a living body is. Every organism has an internal, likely predetermined and genetically imprinted "drive" pushing toward homeostasis. A reward system is necessary to tell the organism, for example, that it is doing right, that it is achieving a good level of integration, that it is heading on along the right track, the track leading through multiple stages (in a sense, this is life …) to the achievement of higher and more complex levels of integration. Higher integration means greater flexibility, which in turn means fitness, better adaptation to changing environments, better chances to pass over genes, and the like. In healthy individuals, the architecture of their goal-state hierarchy expresses their individual "logic of survival" (Damasio, 1999).

Therefore, the reward system can be conceived of as generating a new, non-symbolic kind of representational content, namely, the internal *value* assigned to a certain state. It is important to note that even if nothing like "value" exists in the objective order of things, i.e. if "value" in the normative sense is not a public property observable from a scientific third-person perspective, an internal simu-lation of value within an individual organism can be causally effective—and adap-tive. A *conscious* representation of value, as, for instance, expressed in a subjectively [4] experienced emotional state, has the additional functional advantage of making survival value-related information *globally available* for the selective and flexible control of action, attention, and cognition within a virtual window of presence. It makes this information accessible to many different processing systems at the same time (Baars, 1988; Metzinger, 2003).

Neurally realized goal representations possess a number of interesting features. First, they change the system's functional ontology: it now acts as if something like goals actually were a part of reality, as if maximizing complexity and integration was a good in itself. As a representational system, it makes (or: embodies) an assumption, and this assumption—a representational construction of valence—be-comes causally active. To use a metaphor coined by Francisco Varela, teleology is now "enacted" (Varela *et al.*, 1991).

Second, on the level of phenomenal goal representation, goal representations frequently are transparent. This is to say that the organism "looks through" the

actual representational mechanism and "directly" assigns value to perceived target objects, to specific subject–object relations characterizing certain goal-states, or even to another agent. Because earlier processing stages are not accessible to introspective attention, their content appears as directly and immediately given. On the level of conscious experience, the internal construct of "valence" therefore frequently becomes an objective property perceived in external parts of reality.

However, at least in humans, phenomenally opaque goal representations do exist as well: in these cases we consciously experience them *as* representations. We experience ourselves as beings that actively *construct* goals, and then operate with internal representations of the corresponding goal-states. Conscious goal representations are positioned on a continuum between transparency and opacity, with a strong bias towards transparency (for more on phenomenal transparency, see Metzinger, 2003, Section 3.2.7, 6.4.2).

A third feature makes goal representations interesting: goal representations do not possess truth-conditions, but conditions of satisfaction. Because no such things as goals exist in the objective order of things, a goal representation *cannot* be true or false. However, the goal-*states* as represented can either hold or not hold, or they can hold to various degrees. Therefore, an internally represented goal-state can continuously be matched against sensory input (or even memories, or the results of planning operations).

Fourth, goal representations are intimately linked to an organism's *self-*representation: in standard situations they typically depict an *individual* logic of survival. Only on higher levels of complexity do we find a mentally represented first-person *plural* and socially individuated goal-states. As representations are also dynamical entities, it is best to conceive of the special form of representational content formed by goal-states as an ongoing process that allows an organism to functionally appropriate the fact that certain states of itself, of the world and its social environment are valuable states to herself.

The possession of goal-states, if integrated into a coherent self-model, allows an organism to *own* the logic of its own survival—functionally, representationally, and sometimes even consciously. Goal-states are states the organism wants to bring about, by means of goal representations. Goals—as representational constructs—are fundamental elements of the brain *model* of the world. In particular they are building blocks of *behavioral space* (as represented by the brain). They turn its functional ontology into a teleological ontology, or, in short, into a "teleontology."

On the other hand, the building blocks of behavioral space seem to be representations of possible actions, as seen from no particular perspective, because the empirical evidence (e.g. on mirror neurons, see below) now clearly seems to point to the existence of agent-independent goal detectors. Therefore, a particularly interesting way of analyzing goal representations is as states that portray a successful, completed action from no particular/individual perspective. The theoretical problem to be solved is to explain how precisely these allocentric entities get bound to the conscious perspective of a full-blown agent. But what is an "action"? And what is a "perspective"?

## 4. Actions

Let us start by defining actions. Actions, for some classes of nervous systems, are elementary building blocks of reality, too. Some kinds of organisms have developed agent-detecting modules, and some of them also conceive of *themselves* as agents. They have an extended ontology, because their reality has been considerably enriched. Let us call such systems possessors of an "action ontology." We now have to distinguish movements, behavior, and actions. Bodily movements are simple physical events, and they can be represented accordingly. Behaviors are movements that are goal-directed, i.e. which can meaningfully be described as directed towards a set of satisfaction conditions, but without necessarily being linked to an explicit and conscious representation of such conditions [5]. A dramatic example is presented by a neurological disorder known as anarchic hand syndrome [6]. As simple motor acts, they also do not have a reward-producing component (Rizzolatti *et al.*, 2001, p. 668). In particular, behavior is something that can take place in the absence of conscious self-attribution.

Actions are a specific subset of goal-directed movements: a series of movements that are *functionally integrated* with a currently active goal representation as leading to a reward constitute an action. Therefore, an action is not isomorphic to a particular movement or specific behavioral pattern, because many different movements can constitute the same goal-directed action. What individuates an action is the set of satisfaction conditions defining the representational content of its goal-component as leading to a reward plus the special way in which it is causally linked to the actual event of overt movement generation. In particular, an action results from a *selection process* (which may or may not be conscious) and a representation of the system as a whole as standing in a certain *relation* to a specific goal-state (which is phenomenally represented, e.g. globally available via short-term memory).

The second defining characteristic is that an action in the true sense not only involves an explicit and conscious self-representation, but also a representation of the *perspective* the system now takes onto the world. That is, the selection process may well be unconscious, but it inevitably leads to a more global final stage resulting in a conscious representation of the system as a whole—as having an intention, as initiating and executing its own bodily movements. In other words, on the phenomenal level we always find a corresponding global state in which the system as a whole is itself represented *as* an agent.

This leads us to the third, and most intriguing, aspect of actions in the strong sense we are here proposing: actions are *first-person phenomena*, and they are carried out by a conscious self. It is important to understand that all of this does not necessarily involve reflexive self-consciousness, the possession of concepts, or the mastering of a language: in an animal such as a monkey an attentional and a volitional perspective could suffice to establish the first-person character of actions. In order to be appropriately related to an action goal it is enough to be able to (non-cognitively, subsymbolically) *attend* to it or to (non-cognitively, subsymbolically) *select* a specific motor pattern. At least these are two further assumptions underlying the way in which the brain, as a representational system, typically

decomposes the reality it models into sub-components. We can only understand the special causal linkage between a goal representation and the overt motor output, if we also understand how both are mediated through a specific representational structure, namely, the "phenomenal model of the intentionality relation" (PMIR). This structure will be explored in the next section.

## 5. What is a PMIR? A short representationalist analysis

### 5.1. Four introductory examples

After having clarified the functional structure of the agent–object relationship and its neural underpinnings, we can now focus on the way this relationship can be phenomenally represented. What is the phenomenal model of the intentionality relation (PMIR; see also Metzinger, 2003, Chapter 6)? It is a conscious mental model, and its content is an ongoing, episodic *subject–object relation*. Here are four different examples, in terms of typical phenomenological descriptions of the class of phenomenal states at issue: "I am someone who is currently visually attending to the color of the book in my hands," "I am someone currently grasping the content of the sentence I am reading," "I am someone currently hearing the sound of the refrigerator behind me," "I am someone now deciding to get up and get some more juice."

The first defining characteristic of phenomenal models of the intentionality relation is that they depict a certain *relationship* as currently holding between the system, as transparently represented to itself, and an object component. Phenomenologically, a PMIR creates the experience of a self in the act of knowing or of a self in the act of intending and acting. In the latter case, an object representation is transiently integrated into an ongoing action simulation. This class of phenomenal mental models is particularly rich, because, first, the number of possible object components is almost infinitely large.

Let us now look at our four different examples from a third-person, representationalist stance and see how the object component can vary. In the first situation the object component is formed by a perceptual object, given through the visual and tactile modalities. Put in cornered brackets, the representational content could be described as: [A self in the act of attending to a book in its hands]. In this configuration the object component (the book in the subject's hands) is transparent—that is, the fact that it actually is a *representation* is not introspectively available to the organism as a whole (for more on phenomenal transparency, see Metzinger, 2003, Section 3.2.7). It is plausible to assume that for all animals only capable of constructing a non-cognitive first-person perspective (e.g. monkeys) it is true that they only possess transparent representational content, while human beings are characterized by the important further fact that they can entertain both transparent and opaque object components.

This important further characteristic is demonstrated by the second example, namely, the mental representational content underlying the first-person statement of "I am someone currently grasping the content of the sentence I am reading." The object component, as in this example, can also be formed by a phenomenally *opaque*

kind of conscious content, as in cases of cognitive self-modeling. As explained above, sometimes we experience the content of our own conscious representation as a form of representational content, more often we don't. In this case we do. In the second example the overall representational content contains a phenomenally opaque content element as its object component, because it could be described as [A self in the act of understanding the semantic content of the sentences it is currently reading].

Looking at this second example, we can point out the third defining characteristic for the class of phenomenal mental models of the intentionality relation: the subject component, formed by the currently active self-model, *always* is transparent—although it *may* possess additional opaque partitions. What the second example also shows is how different phenomenal mental models of the intentionality relation make different *types* of relationships between subject and object globally available, e.g. attending, thinking, willing, etc. Of course, all this is not propositional attitude psychology, as the underlying theory of mental representation is of a much more empirically plausible kind. However, it may be helpful to note that, just as in the classical conceptual analysis of propositional attitudes, we still have something like a *content specificator* (the object component), a *person specificator* (the transparent model of the system as a whole), and an *attitude specificator* (the kind of relation between subject and object as represented on the level of conscious experience). This will become more evident, if we consider our next two examples.

In low-level attention, the phenomenally modeled self–object relation is suddenly popping up, as it were, without the accompanying phenomenal quality of agency. Compare our third example: [A self in the act of suddenly noticing the sound of the refrigerator behind him]. Here we have a transparent self-model, a transparent object-model, and the relationship between both of them appears as entirely "unconstructed," as immediately given, as effortlessly and naturally appearing in the conscious model of the world as it were. Of course, there is a complex chain of unconscious neural events preceding the sudden activation of this model. The type of relationship consciously experienced is not [deliberately attending], as in the book case, but [finding yourself forced to automatically attend], as if being driven by the affordances in your environment. What is missing is *attentional agency*, a transparent representation of the process of selecting the object component for attention, as integrated into the self-model.

However, at least in humans, there is not only a cognitive and an attentional first-person perspective, but also a *volitional first-person perspective*. It is a major component of the brain's ontology. It plays a single and unified causal role. And it is at this point that the motor system plays a major role. Interestingly, in systems capable of generating a PMIR, the object component can now also be constituted by the output of ongoing phenomenal simulations, e.g. by the conscious representation of *possible actions*.

From a third-person, representationalist point of view, the fourth and last example can be analyzed as: [A self in the act of selecting a certain possible action, e.g. walking over to the refrigerator and getting another drink]. What typically follows the activation of this kind of representational content is the real *embodiment*

of the action, the running of a complex motor self-simulation, which is now coupled to the effectors. The event just before this stage consisted in phenomenally simulating a number of possible actions and then, as it were, "reaching out" for one of them, making one of them *your own*. This is the moment of integrating the currently active, transparent self-model with an opaque action simulation. "Opaque" means that you know that this, so far, is only a thought, a mentally simulated action. Frequently this new object component may also be rather abstract—an allocentric, as yet self–other-neutral goal representation. By integrating this new object component with the self-model, the brain generates a phenomenal model of the *practical* intentionality relation [7].

Something abstract becomes more concrete. An allocentric representation is integrated into an egocentric frame of reference. The organism—experientially as well as functionally—*appropriates* an action. In other words: a conscious, volitional first-person perspective emerges if a phenomenal model of the system as a whole as standing in relation to a potential *volitional* object component is activated. This model is what we termed a goal representation in Section 3. It assigns a value to a certain state of affairs. In passively desiring a certain object, e.g. the juice to be found in the refrigerator, we have a non-volitional kind of relationship being consciously modeled, and the object component is not formed by a possible action—as in the genuinely volitional perspective—but only by a possible action goal. In actively pursuing this goal-state we additionally appropriate an adequate behavioral pattern.

In short, phenomenal models of the intentionality relation consist of a transparent subject component and varying object components, which can be transparent as well as opaque, transiently being integrated into an overarching, comprehensive representation of the system as standing in a specific relation to a certain part of the world. The overall picture emerging is that of the human self-model continuously integrating the mechanisms of attentional, cognitive and volitional availability against a stable background, which is formed by the transparent representation of the bodily self (Metzinger, 2000, 2003).

### 5.2. The phenomenology of being a volitional subject and a conscious agent

Proceeding from a representationalist level of description to the phenomenological changes inherent in the emergence of a full-blown phenomenal first-person perspective, it is easy to see how it for the first time allows a system to consciously experience itself as being not only a part of the world, but of being fully immersed in it through a dense network of causal, perceptual, cognitive, attentional and agentive relations.

A parallel analysis is possible for the phenomenological properties of *volitional* subjectivity and agency. Conscious volition is generated by integrating abstract goal representations—constituted by self-simulations—into the current model of the phenomenal intentionality relation as object components, in a process of decision or selection. However, let us differentiate a number of cases. If we contemplate a certain action goal, i.e. when we ask ourselves whether we should get up and walk

over to the refrigerator, we experience ourselves as cognitive subjects. This kind of phenomenally represented subject–object relationship can be analyzed in accordance with the model presented in the last section, the only difference being that the object component now is opaque. We know that we take a certain attitude toward a self-generated *representation* of a goal-state.

A completely different situation ensues if we integrate a goal representation into the self-model, thereby making it a part of ourselves by *identifying* with it. Obviously, goal representations and goal hierarchies are important components of self-models that are not based on transient subject–object relations, but on enduring internal reorganizations of the self-model, of its emotional and motivational structure, etc., and which can possibly last for a lifetime.

A *volitional first-person* perspective—the phenomenal experience of practical intentionality—emerges if two conditions are satisfied. First, the object component must be constituted by a particular self-simulatum, by a neural simulation of a concrete behavioral pattern, e.g. like getting up and walking toward the refrigerator. Second, the relationship depicted on the level of conscious experience is one of *currently selecting* this particular behavioral pattern, as simulated.

Again, it is useful to speak about representational identification: the moment following volition, the moment at which concrete bodily behavior actually ensues, is the moment in which the already active motor simulation is *integrated* into the currently active bodily self-model, and thereby causally coupled to the rest of the motor system and the effectors. It is precisely the moment, in which we *identify* with a particular action, transforming it from a possible into an actual pattern of behavior and thereby functionally as well as phenomenologically *embodying* it. Embodiment leads to enacting. Interestingly, the moment of agency seems to be the moment when the phenomenal model of the intentionality relation *collapses*. "Collapse" here means that what previously was a possibility portrayed as distinct from the phenomenal self now becomes a part of it, an object component becomes part of a subject component. We can now describe the experience of being a volitional subject and the experience of being an agent more precisely, using the simple tools already introduced.

Phenomenal volition is a form of phenomenal content, which can be analyzed as *representational* content as follows: [I *myself* ( = the currently active transparent model of the self) am currently ( = the *de-nunc*-character of the overall phenomenal model of the intentionality relation, as integrated into the virtual window of presence) present in a world ( = the transparent, global model of reality currently active) and I am just about to select ( = the type of relation depicted in the phenomenal model of the intentionality relation) a possible way to walk around the chairs toward the refrigerator ( = the object component, constituted by an *opaque* simulation of a possible motor pattern in an egocentric frame of reference)].

The experience of agency follows in the moment in which the internal "distance" created between phenomenal self-representation and phenomenal self-simulation in the previously mentioned structure collapses to zero: I *realize* a possible self, by enacting it. As I experience myself walking around the chairs and toward the refrigerator, proprioceptive and kinesthetic feedback allows me to feel the

degrees to which I have already *identified* with the sequence of bodily movements I have selected in the previous moment.

Remember that transparent representations are precisely those representations the existence of whose content we cannot doubt. They are those we experience as real, whereas opaque representations are those which we experience as thoughts, as imagination, or as hallucinations. *Realizing* a simulated self means devising a strategy of making it the content of a transparent self-model, of a self that really exists—on the level of phenomenal experience.

Ongoing agency, the conscious experience of sustained executive control, can therefore be representationally analyzed according to the following pattern: [I myself (the content of the transparent self-model) am currently ( = the *de-nunc* character of the phenomenal model of the intentionality relationship as integrated into the virtual window of presence) present in a world ( = the transparent, global model of reality) and I am currently experiencing myself as carrying out ( = continuously integrating into the transparent self-model) an action which I have previously imagined and selected (the opaque self-simulation forming the object component, which is now step by step assimilated into the *subject component*)].

Of course, there are all sorts of functional and representational complications, e.g. if the proprioceptive and kinesthetic feedback integrated into the internal model of the body does not match with the forward model still held active in working memory.

In any case, it is interesting to see how agency conceived of as executive consciousness (*Vollzugsbewusstsein*, in the sense of Jaspers) can be analyzed as an ongoing representational dynamics *collapsing* a phenomenal model of the practical intentionality relationship into a new transparent self-model. Again, as the whole structure is embedded into a virtual window of presence, the transparent, intranscendable experiential state for the system itself is one of being a full-blown volitional subject, currently being present in a world, and acting in it.

Please note how the PMIR has a phenomenally experienced *direction*: PMIRs are like arrows pointing from self-model to object component. As soon as one has understood the arrow-like nature of the PMIR, two special cases can be much more clearly described. First, the arrow can point not only outwards, but also *downwards* (phenomenologically speaking: *inwards*). In cases where the object component is formed by a first-order PMIR itself (as in attending to or consciously thinking about oneself *as attending, acting, or thinking*) the second-order PMIR *internally* models a system–system relationship instead of a system–object relationship.

Second, in consciously experienced *social* cognition the object component can now be either formed by a phenomenal model of another agent or an arrow *in* the other agent's head (as in observing another human being observing another human being). Such ideas are appealing, because they show how the relevant representational domain is an open domain. In principle, many layers of complexity and intersubjective metacognition can be added through a process of social/psychological evolution. As the elementary representational building block, the PMIR, gets richer and more abstract, an ascending and cognitively continuous development from the simple portrayal of body-centered subject–object relations to full-blown self–other

modeling becomes conceivable. What changes from monkeys to humans is only the complexity of the self-modeling process.

## 6. Why intentionality should be phenomenalized

The philosophical step just taken consists in *phenomenalizing intentionality*. Phenomenalizing intentionality, we would like to submit, may be a necessary detour, an indispensable first step in the project of *naturalizing* intentionality *tout court.* Meaning and the conscious experience of meaningfulness have to be separated. Generally speaking, mental representations possess two kinds of content: phenomenal content and intentional content. Phenomenal content locally supervenes: once all internal and contemporaneous properties of your nervous system are fixed, the qualitative character of your conscious experience is fixed as well. However, what is not yet determined is if this experiential character also goes along with actual *knowledge about the world*, with intentionality in terms of a biological information-processing system generating states possessing representational—or *intentional*—content. Intentional content, in many cases, is determined by external and non-local factors. Phenomenal content, on the other hand, is what stays invariant regardless of whether a perception is veridical or a hallucination.

However, what we want to point out is how intentionality (on a pre-rational level, probably starting with the sensorimotor system and early levels of attentional processing) is *itself* depicted on the level of phenomenal content. And it is precisely this kind of conscious content, which has guided theoreticians for centuries in developing their own, now classic theories of intentionality. Due to the principle of local supervenience (see above), it has today become highly plausible that *this* aspect of intentionality can be naturalized. The *phenomenal* experience of being an intentional agent, of being a perceiving, attending and cognizing subject can be naturalized. Of course, this does in no way preclude the possibility that intentional content *as such* can never, and maybe even for principled reasons, be naturalized.

But getting the first obstacle out of the way may greatly help in gaining a fresh access to intentionality *as such*, because it frees us from the burden of false intuitions generated by our own transparent model of reality and because it helps us to set aside the issue of how we come to consciously *experience* our mental states as meaningful and directed towards an object component. We can separate the issue of consciously experienced intentionality from the more general problem of how something like representational content could evolve in the minds of human beings and other animals *at all*. If we succeed in anchoring our concept of the PMIR on the functional and neurobiological levels of description, then the notion could even survive a dynamicist revolution. If some day it turns out that, strictly speaking, something like mental content does not even exist (because it *never* existed in the first place), if what is now the most important level of analysis—the "representational stance"—should eventually be abandoned by a science of the mind advanced beyond anything imaginable today, then we would still possess a theory about how it was possible, and necessary, for beings like ourselves to consciously

experience themselves as possessing "contentful" intentional states as described by classical theories of mind.

It is interesting to note how the genuinely philosophical concept of a conscious model of the intentionality relationship (Metzinger, 1993, 2000, 2003) currently surfaces at a number of places in the cognitive neurosciences. Delacour (1997, p. 138), in an excellent review of current ideas about possible neural correlates of conscious experience, explicitly introduces the notion of an "intentionality-modeling structure." LaBerge (1997, pp. 150, 172) points out how important an understanding of the self-representational component present in attentional processing will have to be for a full-blown theory of conscious attention. Craik *et al.* (1999, p. 26) point out how episodic memory, of course, is a process of *re*constructing what was here termed a PMIR, because one necessary constituent of memory retrieval is not simply the simulation of a past event, but an association of this simulation with a *self*-representation. Obviously, building an autobiographic memory is a process of self-related encoding, and conscious, episodic memory retrieval is a process necessarily involving the self-model, because reactivating a PMIR inevitably means reactivating a PSM. Most notable, of course, is Damasio's conception of a "juxtaposition" of self and object (see Damasio & Damasio, 1996a, p. 172; 1996b, p. 24) and the general framework of a fully embodied "self in the act of knowing" (Damasio, 1994, 1999).

Note that the theory is mute about the question if anything like "real" intentionality does exist. Of course, a highly interesting speculation is that philosophical models of the intentionality of the mental have ultimately resulted from a naïvely realistic interpretation of the process of visual attention, of the phenomenal self directing its gaze at a visual object, thereby making it more salient, and simply elevating this interpretation to the level of epistemology. The concept of the intentionality of the mental may simply be a mistaken attempt to theoretically model epistemic relations in accordance with the consciously experienced model of the intentionality relation. Pursuing this point is outside the scope of the present article. But let us at least briefly point to two interesting issues, which are of considerable philosophical interest.

Of particular interest is the fact that the brain models the relationship between subject and object as an asymmetric relationship. It is the consciously experienced "arrow of intentionality," paradigmatically experienced in having the feeling of "projecting" visual attention outwards, as it were, or in attentionally "tracking" objects in the environment. *Intendere arcum*, to bend the bow of the mind and point the arrow of knowledge toward parts of the world is an intuitively plausible and popular philosophical metaphor, in particular in combination with the idea of "direct," magical intentionality.

We can now understand why such an idea will strike beings like us as intuitively plausible: it is *phenomenally* possible, because there is a directly corresponding structural element in our conscious model of reality. Many theoretical models of the representational relationship are implicitly oriented at the phenomenal experience of visual attention, of the *directedness* inherent in the phenomenal model of the intentionality relation. Frequently, the theoretical model we design about ourselves as

cognitive agents is one of organisms, which, *ad libitum*, direct the beam of their "epistemic flashlight" at parts of the world or their own internal lives, as beings, which generate the representational relation *as subjects of experience*. This can lead to the kind of fallacy, which Dennett (cf. 1991, p. 333) has described as "Cartesian materialism" [8].

A related hypothesis is that philosophical theorizing about the intentionality relation has generally been influenced by that aspect of our phenomenal model of reality, which is generated by our strongest sensory modality: if the process of mental representation in general is modeled in accordance with our dominant sensory modality (namely, vision), we will automatically generate *distal objects*—just as we do in our transparent, visual model of reality. If the object component of a PMIR is of an opaque nature, as in genuinely cognitive contents or in goal representation, a philosophical interpretation of these mental contents as *non-physical*, "intentionally inexistent" objects in the sense of Brentano (1874/1973) becomes inevitable.

## 7. A teleofunctionalist analysis of the PMIR

We can now formulate what, from a teleofunctionalist perspective, the advantage of possessing a phenomenal first-person perspective actually consists in. Phenomenal mental models are instruments used to make a certain subset of information currently active in the system globally available for the control of action, for focal attention and for cognitive processing. A phenomenal model of transient subject–object relations makes an enormous amount of new information available *for* the system: all information related to the fact that it is currently perturbed by perceptual objects, that certain cognitive states are currently occurring in itself, e.g. to the fact that certain abstract goal representations are currently active, that there are a number of concrete self-simulations connecting the current system-state with the state the system *would* have if this goal-state would be realized; allowing for selective behavior and the information that it is a system capable of manipulating its own sensory input, e.g. by turning its head and directing its gaze to a specific visual object. A first-person perspective allows a system to conceive of itself as being part of an independent objective order, while at the same time being anchored in it and able to act *on* it as a subject (e.g. Grush, 2000).

Let us now mention one particular application of this representational principle, which may turn out to be of highest relevance. Once a system is capable of representing transient subject–object relations in a globally available manner it becomes possible for the object component in the underlying representational structure to be formed by the *intentions of other beings*. Once again the brain's ontology is expanded, and, as it will become clear in the next section, the motor system plays a crucial role in this functional expansion: a phenomenal first-person perspective allows for the mental representation of a phenomenal *second-person* perspective. The PMIR is what builds the bridge into the social dimension. Once a rudimentary subjective perspective has been established with the help of the motor system, *inter*subjectivity can follow. The dynamics of low-level intersubjectivity then helps to further develop, enrich, and stabilize the individual first-person perspective

in each participating agent. If a functional mechanism for discovering and phenomenally representing the unobservable goal-states of conspecifics is in place, the observed behavior of *other* systems in the organism's environment can lead to an activation of a goal representation, which in turn can be represented as belonging to someone *else*. They would then become a part of the organism's reality, leading to a change in its behavioral profile. As we will see, it is empirically plausible that such a mechanism is actually in place in human beings. Therefore, representations of the intentions of external agents can now become the object component of the phenomenal model of the intentionality relation as well.

Behavior-reading is transformed into mind-reading, because the representational tools for action representation allow for action *simulation* as well (see Gallese, 2003; Gallese & Goldman, 1998), including the simulation of goal-states (i.e. of successfully achieved subject–object relations in *other* agents). If this happens on the level of conscious experience, a completely new and highly interesting form of information becomes globally available for the system: the information of actually *standing in certain relations to the goals of other conspecifics*. We would claim that it is precisely the conscious availability of this type of information, which turned human beings from acting, attending and thinking selves into social subjects. If the fact that your are constantly not only standing in perceptual and behavioral relations to your environment, but that you are frequently realizing *subject–subject relationships* becomes globally available, it also becomes available for cognition.

This, in turn, will allow those systems capable of concept formation to mentally model social relations from a third-person perspective. Such beings can mentally represent social relationships between other individuals depicted as intentional agents, even if they are not involved themselves. We will not pursue this point at length here, but it is obvious how this representational ability is of high relevance for social cognition and the pooling of cognitive resources. Social cognition, empathy, and cognition in general, need a minimal level of complexity in the brain's functional ontology, and in human beings and primates the premotor cortex seems to be substantial part of the physical substrate underpinning this level.

## 8. What is social action *recognition*?

We have so far developed a model of a "neurally realized action ontology," mostly considering the owner of this ontology in isolation. However, primates, and particularly human beings are social animals whose cognitive development capitalizes upon the interaction with other conspecifics (adults, siblings, etc.). During social interactions we overtly manifest our inner intentions, dispositions and thoughts by means of overt behavior. We reciprocate this by trying to figure out what are the intentions, dispositions and thoughts of others when witnessing their behavior. Detecting another agent's intentions or other inner states helps to anticipate this agent's future actions, which may be cooperative, non-cooperative, or even threatening. Accurate understanding and anticipation enable the observer to adjust his responses appropriately.

Some recent neuroscientific results seem to suggest that a common neural

representation underpins the specification of action goals, independently of whose goals are to be specified. A series of single neuron recording experiments led to the discovery in two reciprocally connected sectors of the monkey brain, premotor area F5 and parietal area 7b, of a particular set of neurons, activated both during the execution of purposeful, goal-related hand actions, such as grasping, holding or manipulating objects, and during the observation of similar actions performed by another individual. These neurons were designated as "mirror neurons" (Gallese, 2000a, 2001; Gallese *et al.*, 1996, 2002; Rizzolatti *et al.*, 1996a; see also Fogassi & Gallese, 2002; Rizzolatti *et al.*, 2000, 2002).

Mirror neurons require, in order to be activated by visual stimuli, an interaction between the agent (be it a human being or a monkey) and its object. The visual presentation of objects alone does not evoke any response. Similarly ineffective or very little effective in driving the neurons response are actions that, although achieving the same goal and looking similar to those performed by the experimenter's hand, are made with tools such as pliers or pincers. Actions having emotional content, such as threatening gestures, are also ineffective. The general picture emerging is that the sensorimotor integration process supported by the fronto-parietal F5-PF mirror matching system instantiates an "internal copy" of actions utilized not only to generate and control goal-related behaviors, but also to provide—at a pre-conceptual and pre-linguistic level—a meaningful account of behaviors performed by other individuals.

Several studies that used different methodologies have demonstrated also in humans the existence of a similar mirror system, matching action observation and execution (see Buccino *et al.*, 2001; Cochin *et al.*, 1998; Decety *et al.*, 1997; Fadiga *et al.*, 1995; Grafton *et al.*, 1996; Hari *et al.*, 1998; Iacoboni *et al.*, 1999; Rizzolatti *et al.*, 1996b).

In particular, it is interesting to note that brain imaging experiments in humans have shown that during action observation there is a strong activation of both premotor and parietal areas (Buccino *et al.*, 2001; Decety & Grèzes, 1999; Decety *et al.*, 1997; Grafton *et al.*, 1996; Iacoboni *et al.*, 1999; Rizzolatti *et al.*, 1996b), which very likely are the human homologue of the monkey areas in which mirror neurons were found. Furthermore, in humans both lesions in premotor area 44 (the area homologue of F5; see Rizzolatti & Arbib, 1998) and in the inferior parietal lobe (in which area 40 is likely the homologue of monkey's area PF) produce deficits in action recognition (Bell, 1994; Brain, 1961; Duffy & Watkins, 1984; Gainotti & Lemmo, 1976; Heilman & Rothi, 1993; Heilman *et al.*, 1982). This directly demonstrates how representational reduction leads to functional deficits. The functional ontology of these patients has been reduced in its capacity to be expanded, because the necessary tools—action and action representation—are not available to the system any more. Their model of reality has been impoverished.

Let us investigate more closely what relationship could exist between planning an action and understanding the action of others. When a given action is planned, its expected motor consequences are forecast. This means that when we are going to execute a given action we can also predict its consequences. This prediction is likely the computational result of the action model. Through a process of

"equivalence" between what is acted and what is perceived, this information can also be used to predict the consequences of actions performed by others. This equivalence—underpinned by the activity of mirror neurons—is made possible by the fact that both predictions (of our actions and of others' actions) are simulation (modeling) processes. The same functional logic that presides over self-modeling and PMIR construction is employed also to model the behavior of others: to perceive an action is equivalent to internally simulating it. This enables the observer to use her/his own resources to penetrate the world of the other by means of an *implicit*, *automatic*, and *unconscious* process of motor simulation. Such simulation process automatically establishes a direct link between agent and observer (for a discussion of mirror neurons, empathy and simulation theory, see Gallese, 2001, 2003; Gallese & Goldman, 1998). At a later stage, on the level of phenomenal representation, it generates an isomorphy between the object components of their PMIRs. This isomorphy carries information, which can be functionally exploited.

Perhaps even more intriguing are recent data from monkey experiments conducted by Umiltà *et al.* (2001), in which F5 mirror neurons were tested when the monkey could see the entire action (e.g. a hand grasping action), or when the same action was presented but its final critical part, that is the hand–object interaction, hidden. In the hidden condition the monkey knew that the target object was present behind the occluder. The results showed that more than half of the recorded neurons responded also in the hidden condition. These data indicate that, as humans, also monkeys can infer the goal of an action even when the visual information about it is incomplete. This inference appears to be mediated by the activity of motor neurons coding the (goal-) end state of the action.

A representational analysis of such neural mechanisms clearly shows that the object component of an other-agent PMIR simulated by F5 mirror neurons does not have to be visible in order for the full-blown neural response to occur. This proves that mirror neurons code not object presence, but rather the *relational fact* that a certain external agent is currently *directed at an object component*. They code the existence of a PMIR in *another* agent, and interestingly they do so in virtue of being a central part of the system that, in other situations, constructs an internal PMIR in the monkey herself.

To have a PMIR means to have a phenomenal self-model. To be self-conscious does *not* imply having language, concepts, being able to mentally form a concept of *oneself*, etc. Body image and visceral feelings are enough. Because the monkey's motor system allows for prediction of actions/occluded target objects, neurons responsive to the observation of goal-related behaviors might actually be triggered not by a visual representation alone, but by an "embedded motor schema." We posit that the representational deep structure of this schema is what later evolved into the full-blown PMIR in human beings.

Again, please note that the notion of a PMIR as here introduced does *not* yet imply the possession of language and concepts, of being able to mentally form a concept of oneself *as oneself*, of consciously experiencing the selection process for goal-states, etc. An elementary self-model in terms of body image and visceral feelings plus the feelings plus the existence of a low-level attentional mechanism is

quite enough to establish the basic representation of a dynamic subject–object relation.

Actions in the external world can be experienced as such, recognized and understood only in virtue of a shared action ontology. An action ontology can only be shared by two systems, if there is a sufficient degree of functional overlap between them, if, as we put it at the beginning of this article, they decompose target space in similar ways. The cognitive development of social competence then capitalizes upon such a shared ontology to trigger the timely onset of behaviors such as gaze following, shared attention, and detection of intentions, which will eventually give origin to a full-blown capacity to entertain mental accounts of the behavior and goal-states of other agents. What makes humans special is the fact that their functional ontology is much richer in *socially individuated* goal representations and that their model of reality is not only richer and much more flexible, but that they can actively *expand* their own functional ontology by mentally ascribing distal goals to conspecifics.

Interestingly, in social cognition, what acting and observing system frequently share is the same *intentional object*: both systems internally activate a PMIR sharing the object component. The agent is directed at a certain goal-state, in a *practical* model of the intentionality relation. His internal goal representation has satisfaction conditions. The observing system, extracts the goal-state by automatically and unconsciously simulating the agent, thereby also *emulating* it, because it aims not only internally representing the observable behavior of the target system, but also an abstract and unobservable property of this system, namely, the fact that is directed towards certain satisfaction conditions. The observer is directed at the very same goal-state (via behavioral simulation/emulation), in a *theoretical* model of the intentionality relation. But his internal goal representation of the other agent has truth-conditions: his mind-reading is either true or false.

Therefore, mirror neurons in the monkey brain can perhaps be interpreted as the dawning of what the equivalent matching systems in our human brains are the fully-developed realization of: a fundamental and mostly unconscious representational structure capable to build a *shared* action ontology. In closing, let us give a centrally important example for a high-level utilization of this structure. Two or more human beings can now *at the same time* activate cognitive PMIRs mutually pointing to each other under a representation as rational individuals. And the correlated nature of these two mental events, their mutuality and interdependence, can itself be represented on the level of global availability. Beings like ourselves are therefore able to mutually *acknowledge each other as persons*, and to consciously experience this very fact at the same time.

## 9. Conclusion

To have an ontology is to interpret a world. In this paper we have argued that the brain, viewed as a representational system aimed at interpreting the world, possesses an ontology too: it decomposes target space in a way that exhibits certain primitives and invariances, which in turn are functionally significant. We have then focused on

the contribution the motor system makes to this process. As it turns out, the motor system constructs goals, actions, and intending selves as basic constituents of the world it interprets. It does so by assigning a single, unified causal role to them. Empirical evidence now clearly shows how the brain actually codes movements and action goals in terms of multimodal models of organism–object relations.

We have pointed out how this process, under a representationalist analysis, can be interpreted as an internal representation of the intentionality relation itself, and how such a representation, once it is in place, can later function as the building block for social cognition and for a more complex, *consciously* experienced representation of first-person perspective as well. The motor system may therefore play a decisive role in understanding how the functional ontology of the human brain could be extended into the social domain, how a phenomenal self and the goals of other human beings could become a causally active part of *our reality*.

## Acknowledgements

## Notes

[1] Representationalist approaches always carry the burden of not creating a new version of the homunculus problem: personal-level predicates or what philosophers sometimes call "intentionalist idioms" have to be avoided on all subpersonal levels of description. There is no little man in the head interpreting representations, and, for the approach sketched here, the phenomenal self is something that has no transcendental subject or conscious agent "behind it."

[2] We will treat any system as a "physical" system, which possesses a true description under our currently best physical theories. Brains are such systems. We also want to indicate that eventually the physical level of description may turn out to be relevant to the project we are describing, because the relevant theory of dynamical self-organization may span many levels of description through certain bridge-laws, and because it may be necessary to describe the system's functional interdependence with the environment.

[3] By "objective" world we here mean reality as grasped under third-person, theoretical representations as typically found in successful theories within the natural sciences. The "subjective" world would correspondingly be the world as represented under conscious, first-person models of reality generated by individual organisms. As such, it has no observable properties under intersubjective and supra-individual types of representation.

[4] By "subjective" we mean the domain of phenomenal representation, throughout this paper.

[5] We will treat an *explicit* representation as one in which changes in the representandum invariably lead to a change on the content level of the respective medium. *Implicit* representation will only change functional properties of the medium—for instance, by changing synaptic weights and moving a connectionist system to another position in weight-space. Conscious content will generally be explicit content in that it is globally available and directly covaries with its object. This does, of course, not mean that it has to be linguistic or *conceptually* explicit content.

[6] Anarchic hand syndrome (first described as alien hand syndrome by Goldstein, 1908; Sweet, 1941; terminologically introduced by Brion & Jedynak, 1972; Goldberg *et al.*, 1981; for an important recent differentiation, see Marchetti & Della Sala, 1998) is characterized by a global experiential

state in which the patient typically is well aware of complex, observable movements carried out by the non-dominant hand, while at the same time experiencing no corresponding volitional acts. Subjectively (as well as functionally) the arm is "out of control," with a sense of intermanual conflict. For example, a patient may pick up a pencil and begin scribbling with the right hand, but react with dismay when her attention is directed to this fact. Then she will immediately withdraw the pencil, pull the right hand to her side with the left hand and indicate that she had not *herself* initiated the original action (Goldberg *et al.*, 1981, p. 684). Not only is there no explicit goal representation, and no phenomenal sense of volitional ownership, but we even find an attribution of quasi-personhood in terms of agency and autonomy of the patient to one of her body parts.

[7]   In classical conceptions of intentionality there are two basic types of intentionality, *theoretical* and *practical* intentionality. Theoretical intentionality is aimed at gaining knowledge about the world. In terms of the classic propositional attitude framework, theoretical intentionality is realized by cognitive attitudes, possessing content specificators exhibiting truth values—as in thinking, believing, etc. Practical intentionality, however, is realized by the class of volitive attitudes, e.g. as in wishing, desiring, etc. In these cases the content specificator only possesses *fulfillment conditions* and it points at a certain action goal, in terms of a changed state of affairs in the world. We propose that a consciously perceived volitional first-person perspective emerges precisely if the second class of object components and relations is phenomenally modeled in accordance with the current background assumptions.

[8]   As Dennett has pointed out, many of the different forms of Cartesian materialism, the assumption of a final inner stage, can also be generated in the context of representationalist theories of mind by mistakenly transporting what he called the "intentional stance" (Dennett, 1987a) into the system (cf. Dennett, 1991, p. 458). The model here proposed, of course, does not make this mistake; it is much more closely related to the idea of a "second-order intentional system" (Dennett, 1987b), a system that applies the intentional stance to itself—but in a phenomenally transparent manner.

# References

BAARS, B.J. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

BELL, B.D. (1994). Pantomime recognition impairment in aphasia: an analysis of error types. *Brain and Language*, *47*, 269–278.

BRAIN, W.R. (1961). *Speech disorders: aphasia, apraxia and agnosia*. Washington, DC: Butterworth.

BRENTANO, F. (1874/1973). *Psychology from an empirical standpoint*, C. ANTOS, D.B. RANCURELLO & L. MCALISTER (Trans.). London: Routledge & Kegan Paul.

BRION, S. & JEDYNAK, C.-P. (1972). Troubles du transfert interhémisphérique (callosal disconnection). A propos de trois observations de tumeurs du corps calleux. Le signe de la main étrangère. *Revue Neurologique (Paris)*, *126*, 257–266.

BUCCINO, G., BINKOFSKI, F., FINK, G.R., FADIGA, L., FOGASSI, L., GALLESE, V., SEITZ, R.J., ZILLES, K., RIZZOLATTI, G. & FREUND, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience*, *1*, 400–404.

CHAO, L.L. & MARTIN, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, *605*, 478–484.

COCHIN, S., BARTHELEMY, C., LEJEUNE, B., ROUX, S. & MARTINEAU, J. (1998). Perception of motion and qEEG activity in human adults. *Electroenc Clinical Neurophysiology*, *107*, 287–295.

CRAIK, F.I.M., MOROZ, T.M., MOSCOVITCH, M., STUSS, D.T., WINOCUR, G., TULVING, E. & KAPUR, S. (1999). In search of the self: a positron emission tomography study. *Psychological Science*, *10*, 26–34.

DAMASIO, A.R. (1994). *Descartes' error*. New York: Putnam.

DAMASIO, A.R. (1999). *The feeling of what happens: body and emotion in the making of consciousness*. New York: Harcourt Brace.

DAMASIO, A.R. & DAMASIO, H. (1996a). Images and subjectivity: neurobiological trials and tribulations. In R.N. MCCAULEY (Ed.) *The Churchlands and their critics*. Cambridge, MA: Blackwell.

DAMASIO, A.R. & DAMASIO, H. (1996b). Making images and creating subjectivity. In R. LLINÁS & P.S. CHURCHLAND (Eds), *The mind–brain continuum*. Cambridge, MA: MIT Press.

DECETY, J. & GRÈZES, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences*, *3*, 172–178.

DECETY, J., GREZES, J., COSTES, N., PERANI, D., JEANNEROD, M., PROCYK, E., GRASSI, F. & FAZIO, F. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain*, *120*, 1763–1777.

DELACOUR, J. (1997). Neurobiology of consciousness: an overview. *Behavioural Brain Research*, *85*, 127–141.

DENNETT, D.C. (1987a). *The intentional stance*. Cambridge, MA: MIT Press.

DENNETT, D.C. (1987b). Intentional systems in cognitive ethology: the "Panglossian Paradigm" defended. *Brain and Behavioral Sciences*, *6*, 343–390.

DENNETT, D.C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.

DUFFY, J.R. & WATKINS, L.B. (1984). The effect of response choice relatedness on pantomime and verbal recognition ability in aphasic patients. *Brain and Language*, *21*, 291–306.

FADIGA, L., FOGASSI, L., PAVESI, G. & RIZZOLATTI, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology*, *73*, 2608–2611.

FOGASSI, L. & GALLESE, V. (2002). The neural correlates of action understanding in non human primates. In M. STAMENOV & V. GALLESE (Eds) *Mirror neurons and the evolution of brain and language* (pp. 13–36). Philadelphia, PA: John Benjamins.

GAINOTTI, G. & LEMMO, M.S. (1976). Comprehension of symbolic gestures in aphasia. *Brain and Language*, *3*, 451–460.

GALLESE, V. (2000a). The acting subject: towards the neural basis of social cognition. In T. METZINGER (Ed.) *Neural correlates of consciousness: empirical and conceptual questions* (pp. 325–333). Cambridge, MA: MIT Press.

GALLESE, V. (2000b). The inner sense of action: agency and motor representations. *Journal of Consciousness Studies*, *7*, 23–40.

GALLESE, V. (2001). The "shared manifold" hypothesis: from mirror neurons to empathy. *Journal of Consciousness Studies*, *8*, 33–50.

GALLESE, V. (2003). The manifold nature of interpersonal relations: the quest for a common mechanism. *Philosophical Transactions of the Royal Society of London*, *358*, 517–528.

GALLESE, V. & GOLDMAN, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *12*, 493–501.

GALLESE, V., FADIGA, L., FOGASSI, L. & RIZZOLATTI, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.

GALLESE, V., FOGASSI, L., FADIGA, L. & RIZZOLATTI, G. (2002). Action representation and the inferior parietal lobule. In W. PRINZ & B. HOMMEL (Eds) *Attention and performance XIX: common mechanisms in perception and action* (pp. 334–355). Oxford: Oxford University Press.

GENTILUCCI, M., FOGASSI, L., LUPPINO, G., MATELLI, M., CAMARDA, R. & RIZZOLATTI, G. (1988). Functional organization of inferior area 6 in the macaque monkey. I. Somatotopy and the control of proximal movements. *Experimental Brain Research*, *71*, 475–490.

GOLDBERG, G., MAYER, N.H. & TOGLIA, U.T. (1981). Medial frontal cortex infarction and the alien hand sign. *Archives of Neurology*, *38*, 683–686.

GOLDSTEIN, K. (1908). Zur Lehre der motorischen Apraxie. *Journal für Psychologie und Neurologie*, *11*, 169–187.

GRAFTON, S.T., ARBIB, M.A., FADIGA, L. & RIZZOLATTI, G. (1996). Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Experimental Brain Research*, *112*, 103–111.

GRUSH, R. (2000). Self, world and space: the meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind*, *1*, 59–92.

HARI, R., FORSS, N., AVIKAINEN, S., KIRVESKARI, S., SALENIUS, S. & RIZZOLATTI, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences USA*, *95*, 15,061–15,065.

HEILMAN, K.M. & ROTHI, L.J. (1993). Apraxia. In K.M. HEILMAN & E. VALENSTEIN (Eds) *Clinical neuropsychology* (pp. 141–163). New York: Oxford University Press.

HEILMAN, K.M., ROTHI, L.J. & VALENSTEIN, E. (1982). Two forms of ideomotor apraxia. *Neurology*, *32*, 342–346.

HEPP-REYMOND, M.C., HÜSLER, E.J., MAIER, M.A. & QI, H.-X. (1994). Force-related neuronal activity in two regions of the primate ventral premotor cortex. *Canadian Journal of Physiology and Pharmacology*, *72*, 571–579.

IACOBONI, M., WOODS, R.P., BRASS, M., BEKKERING, H., MAZZIOTTA, J.C. & RIZZOLATTI, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–2528.

JEANNEROD, M., ARBIB, M.A., RIZZOLATTI, G. & SAKATA, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neuroscience*, *18*, 314–320.

KURATA, K. & TANJI, J. (1986). Premotor cortex neurons in macaques: activity before distal and proximal forelimb movements. *Journal of Neuroscience*, *6*, 403–411.

LABERGE, D. (1997). Attention, awareness, and the triangular circuit. *Consciousness and Cognition*, *6*, 149–181.

MARCHETTI, C. & DELLA SALA, S. (1998). Disentangling the alien and the anarchic hand. *Cognitive Neuropsychiatry*, *3*, 191–207.

MARTIN, A., WIGGS, C.L., UNGERLEIDER, L.G. & HAXBY, J.V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*, 649–652.

MATELLI, M., LUPPINO, G. & RIZZOLATTI, G. (1985). Patterns of cytochrome oxidase activity in the frontal agranular cortex of the macaque monkey. *Behavioral Brain Research*, *18*, 125–137.

METZINGER, T. (1993). *Subjekt und Selbstmodell*. Paderborn: Mentis.

METZINGER, T. (2000). The *subjectivity* of subjective experience: a representationalist analysis of the first-person perspective. In T. METZINGER (Ed.) *Neural correlates of consciousness: empirical and conceptual questions*. Cambridge, MA: MIT Press.

METZINGER, T. (2003). *Being no one: the self-model theory of subjectivity*. Cambridge, MA: MIT Press.

MURATA, A., FADIGA, L., FOGASSI, L., GALLESE, V., RAOS, V. & RIZZOLATTI, G. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology*, *78*, 2226–2230.

PERANI, D., CAPPA, S.F., BETTINARDI, V., BRESSI, S., GORNO-TEMPINI, M., MATARRESE, M. & FAZIO, F. (1995). Different neural systems for recognition of animals and man-made tools. *Neuroreport*, *6*, 1637–1641.

RIZZOLATTI, G. & ARBIB, M.A. (1998). Language within our grasp. *Trends in Neuroscience*, *21*, 188–194.

RIZZOLATTI, G., SCANDOLARA, C., MATELLI, M. & GENTILUCCI, M. (1981). Afferent properties of periarcuate neurons in macaque monkeys. II. Visual responses. *Behavioral Brain Research*, *2*, 147–163.

RIZZOLATTI, G., CAMARDA, R., FOGASSI, L., GENTILUCCI, M., LUPPINO, G. & MATELLI, M. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Experimental Brain Research*, *71*, 491–507.

RIZZOLATTI, G., FADIGA, L., FOGASSI, L. & GALLESE, V. (1996a). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

RIZZOLATTI, G., FADIGA, L., MATELLI, M., BETTINARDI, V., PAULESU, E., PERANI, D. & FAZIO, F. (1996b). Localization of grasp representation in humans by PET: 1. Observation versus execution. *Experimental Brain Research*, *111*, 246–252.

RIZZOLATTI, G., FOGASSI, L. & GALLESE, V. (2000). Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In M.S. GAZZANIGA (Ed.) *The cognitive neurosciences* (pp. 539–552). Cambridge, MA: MIT Press.

RIZZOLATTI, G., FOGASSI, L. & GALLESE, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews in Neuroscience*, *2*, 661–670.

RIZZOLATTI, G., CRAIGHERO, L. & FADIGA, L. (2002). The mirror system in humans. In M. STAMENOV & V. GALLESE (Eds) *Mirror neurons and the evolution of brain and language* (pp. 37–62). Philadelphia, PA: John Benjamins.

SWEET, W.H. (1941). Seeping intracranial anuerysm simulating neoplasm. *Archives of Neurology and Psychiatry*, *194*, 86–104.

UMILTÀ, M.A., KOHLER, E., GALLESE, V., FOGASSI, L., FADIGA, L., KEYSERS, C. & RIZZOLATTI, G. (2001). "I know what you are doing": a neurophysiological study. *Neuron*, *32*, 91–101.

VARELA, F., THOMPSON, E. & ROSCH, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.