THE ROYAL SOCIETY

Published online 4 June 2003

A neuroscientific grasp of concepts: from control to representation

Vittorio Gallese

Dipartimento di Neuroscienze, Sezione di Fisiologia, Università di Parma, Via Volturno 39, I-43100 Parma, Italy (*vittorio.gallese@unipr.it*)

Abstraction denotes the cognitive process by means of which general concepts are formed. The dominant view of abstraction considers it not only as a complex and sophisticated cognitive activity, but also as a distinctive hallmark of mankind. The distinctiveness of abstract thought has indeed been closely related to another feature peculiar to our species: language. Following this perspective, the possibility to entertain conceptual representations is thus precluded to animals devoid of full-blown language. I challenge this view and propose that the representational dynamic of the brain is conceivable as a type of self-organization, in which action plays a crucial part. My aim will be to investigate whether, and to what extent, conceptual knowledge can be attributed to non-linguistic animal species, with particular emphasis on non-human primates. I therefore introduce the notion of semantic content as a type of 'relational specification'. A review of recent neurophysiological data on the neural underpinnings of action end-states in the macaque monkey brain is presented. On the basis of this evidence, I propose that conceptual representations can be conceived as the expression of a coherent internal world model. This model decomposes the 'outer' space inhabited by things in a meaningful way only to the extent that it accords to biologically constrained, embodied invariance. Finally, I discuss how the 'comparative' neuroscientific approach to abstraction proposed here may shed some light on its nature and its evolutionary origin.

Keywords: monkeys; mirror neurons; concepts; naturalization

1. INTRODUCTION

> Human concepts are not random or arbitrary; they are highly structured and limited, because of the limits and structure of the brain, the body, and the world.
>
> (Lakoff & Nuñez 2000, p. 1)

The notion of abstraction has a long philosophical history. Briefly, it denotes the cognitive process by means of which general concepts are formed. According to this view, abstraction requires the omission of each and every particular feature from our knowledge of a given collection of things or state of affairs, while preserving most relevant commonalities, whatever they might be.

The overall still dominant view of abstraction—the process enabling concepts to be formed—considers it as not only a complex and sophisticated cognitive activity, but also a distinctive hallmark of mankind. At first glance, this view is intuitively very plausible. When we think of abstract thought, we immediately connect it with mathematics, the operations of formal logic, and a host of other cognitive activities that seem more than plausible to conceive as totally alien to non-human species (but see Lakoff & Nuñez 2000). The distinctiveness of abstract thought has indeed been closely related to another feature peculiar to our species: language. Following this perspective, the possibility to entertain conceptual representations is thus precluded to animals devoid of full-blown language. I will challenge this view by showing that the absence of linguistic competence does not necessarily also entail the absence of conceptual knowledge. I substantiate this proposal with neuroscientific evidence. Before putting my cards on the table, though, it is worthwhile looking more closely at the mainstream view.

Classic cognitive science conceives concepts as abstract, amodal and arbitrary propositions represented in some 'language of thought' (Fodor 1975, 1987), which, although not necessarily identical to language, shares with it many features: productivity and compositionality, among others. The propositional picture of the mind conveyed by classic cognitivism is that of a functional systemwhose processes can be described in terms of manipulations of informational symbols according to a set of formal syntactic rules (see Fodor 1981; Pylyshyn 1984). Input information is symbolically represented and progressively transformed for perception and motor output. Knowledge is therefore represented in symbolic form. Meaning is referential, since it derives from the posited nomological correspondence between the system of symbols and their corresponding extensions, the objects and events in the world. These latter are classified in terms of classical categories, whose membership is defined in terms of singly necessary and jointly sufficient conditions. Thus, following the line of arguments of classic cognitivism, concepts are symbolic representations in their nature, and as thinking, they can be reduced to computation.

This approach to the mind denotes both the powerful

One contribution of 16 to a Theme Issue 'The abstraction paths: from experience to concept'.

influence of mathematical logic and artificial intelligence, and that of the Chomskian 'universal grammar' theory (Chomsky 1986). This approach appealed to many, hence its interdisciplinary popularity, because it guarantees a 'politically correct' materialism, while preserving the sacred aura of the mind/soul, a mind conceived as totally remote from the trivialities of the life of the body. For the 'concepts-as-computational-symbols' view, the mechanisms enabling the relation between the cognitive agent and the 'real world' are of no relevance for the determination of conceptual representational content, and for the understanding of what this content is and what it stands for (see Fodor 1998). Henceforth, any serious attempt to provide a neuroscientific account of mental representations should be dismissed as useless (see Fodor 2001).

This is no free lunch, however. It is my opinion that neuroscience is, today, in a position to unveil how outrageously high the price to be paid is. Let us see why. This classic view has been challenged from many different perspectives (for an 'experiential' approach to cognition, see Lakoff & Johnson (1980, 1999), Lakoff (1987) and Lakoff & Nunez (2000)). I will limit myself here to emphasizing that if all concepts are supposed to be innate, one condemns oneself to the absurdity of positing that even concepts such as BEAUROCRAT and DOOR-KNOB have to be innate.

Alternatively, if it is conceded that *some* concepts are acquired, the question arises of how the acquisition of those concepts might proceed. The classic cognitivist therefore bears the burden of showing how that question can be answered, while simultaneously refuting the relevance of neuroscience. No plausible answers have so far been provided from the classic cognitivist enclave. Another closely related major problem with this approach consists in its incapacity to provide a consistent account of how symbols can be grounded in their real world referents.

It is out of the scope of this paper, however, to delve deeply into this debate. The issue at hand here will be much more confined and focused. My aim will be to investigate whether, and to what extent, conceptual knowledge can be attributed to non-linguistic animal species, with particular emphasis on non-human primates. To that purpose, I provide a deflationary definition of content that will show how the results of neurophysiological research may indeed provide the bedrock for a plausible naturalization of intentional content. In §§ 1 and 2, I address the notion of 'natural information' and its relevance for the neuroscientific approach. In § 3, I discuss neuroscientific evidence in relation to the issue of how representational content relates to its informational vehicle. This introduces the notion of semantic content as a type of 'relational specification'. At this point, action and the motor system are introduced. In § 4 I introduce and discuss neurophysiological data from our laboratory on the neural underpinnings of action end states in the macaque monkey brain. On the basis of this evidence, I suggest that conceptual representations can be conceived as the expression of a coherent internal world model, which decomposes the space inhabited by things and tokens them in a meaningful way, according to biologically constrained, embodied invariance. In § 5 I discuss how the 'comparative ' neuroscientific approach to abstraction proposed here may shed *some* light on its nature and its evolutionary origin. Some preliminary conclusions are drawn in § 6.

2. NATURALIZING INFORMATION

Concepts constitute a particular kind of representational content. Any attempt to provide a neuroscientific account of representational content therefore also implies an ability to operate a *naturalization* of information, that is, to determine how information embodies content. This will be the focus of this section.

One of the most influential attempts to naturalize information is that proposed by Fred Dretske, within his lifelong broader scope to naturalize intentional content in terms of information theory (1981, 1988, 1995). According to Dretske, a series of distinctions should be drawn: between the informational signal and its meaning, between perceptual and cognitive conceptual information. The information we gather from the environment through sensory channels is equated to analogue-like information, while the content of our cognitive systems is maintained in a digital format. The analogue-to-digital conversion process, conceived as the information-theoretic spelling of generalization, is supposed to narrow down the informational content to produce reliable knowledge. This semantic knowledge, in turn, will form concepts only insofar as it will be part of a belief, thus playing a functional role to guide behaviour (Dretske 1981). The crucial move to connect information to intentional content consists in introducing teleology (see also below), by claiming that it is the *natural function of intentional representations to carry information.*

I would like to explore a different path, namely, that it is the *natural function of natural information to produce intentional representations*, concepts included. Let us start by observing what information is about in biological agents. Each biological agent constantly exchanges information with the environment. This exchange is required to relate to the environment, navigate in it, and to act upon it. If we analyse at the *physical level of description* the relationship between biological agents and 'the world outside', we will find living organisms processing the different epiphanies of energy they are exposed to: electromagnetic, mechanical, chemical energy. Energy interacts with living organisms. It is only by virtue of this interaction that energy can be specified in terms of the 'stimuli' (visual, auditory, somatosensory, etc.) to which every organism is exposed. The result of the interaction between energy and living organisms is that the energy, now 'stimulus', is translated, or better, transduced into a *common informational code*. The receptors of the different sensory modalities are the agents of the transduction process: they convert the different types of energies resulting from organisms–world interactions into the common code of action potentials. Action potentials express the electrochemical excitability of cells, and constitute the code used by the billions of neurons that comprise the central nervous system to 'communicate' with each other. But *sensu stricto*, this is still a code not a language. Information becomes, nevertheless, available to perceive, plan and execute actions.

Thus, we have different informational components. Philosophers and cognitive scientists generally conceive

representations as the *vehicle*, and things to be represented as their *content*. Furthermore, it is commonly argued that it is erroneous to confuse the personal level of description, the level pertaining to mental representations, or propositional attitudes, and the sub-personal level, typical of the objects of neuroscientific investigation.

Perhaps this might work if the aim is to provide a solipsistic account of the possible formal logic of thought. However, I seriously doubt that such a picture *exhaustively* portrays what our mind is, how it works, and where it comes from. If our ultimate goal is to provide a biologically plausible account of the cognitive capacities of *real and situated biological agents*, this dichotomous account appears to be insufficiently apt to do justice to all the elements in play: facts and objects in real and possible worlds; the firing of neurons in the brain; the mental representations that brain activity subsumes.

A coherent and neuroscientifically plausible framework requires all elements to be accommodated within it. My proposal is that this goal can be safely pursued if we consider the 'representational dynamic' of the brain to be *non*-symbolic. This entails consideration of the representational dynamic of the brain as a particular type of self-organization, with body action playing a major role in specifying the informational routines characterizing self-organization (see below).

The non-symbolic qualification is very important, because it provides the opportunity to avoid the circularity and self-referential quality of the symbolic–computational approach (see § 5). It should be noted, *en passant*, that this circularity is indeed well epitomized by the basic failure of the enterprise in which artificial intelligence—heavily inspired by the symbolic–functionalist approach—bravely engaged itself: to reproduce in non-biological media even limited aspects of human cognition. Non-biologically grounded algorithmic accounts of cognition thus square with a hypothetical syntax devoid of any semantics.

The solution I propose is to consider the information processing carried out by the brain of an organism in the larger frame of the interactions between the organism and the environment it is acting upon. Following this perspective, the vehicle/content, sub-personal/personal dichotomies appear as ontological misconceptions, prompted by the mistake of considering information processing in isolation, neglecting *how* and *why* information from the 'world outside' is gathered and processed: to control the interaction between organisms and their worlds. The brain, a brain wired to a body that constantly interacts with the world is, at the same time, the vehicle of information *and* part of its content, the latter being conceived as a way to model organism–environment interactions. My aim is to show how such a deflationary account of representational content may prove useful to tackle these issues from a naturalized and biologically sound perspective.

Here, philosophy is highly relevant and can perhaps offer assistance. Several authors have independently developed naturalistic teleological theories of content (also denoted as 'teleosemantics'), which couple representational content with the notion of purpose (Stampe 1977; Evans 1982; Dretske 1988; Millikan 1984, 1993; Papineau 1987). Teleological theories of mental content were triggered by the need to solve the paradox pointed out by Brentano (1973), when defining the 'aboutness' of intentionality. The paradox is posed to any materialist naturalistic theory of the mind by the possibility to entertain representations of non-existent things, the famous possible worlds full of unicorns, Red Riding Hoods, James Bonds, etc.

According to teleosemantics, positing that any mental content is determined by whatever it is the purpose of the mental state to represent reportedly solves the paradox. Teleosemantics provides, therefore, a teleofunctional account of what produces the semantic content of mental representations.

I will neither commit myself to teleosemantics nor discuss at length its versions, articulations and the criticisms it has engendered. Rather, I will try to exploit from my peculiar neuroscientific standpoint some of the suggestions emanating from this approach to score my point.

'Proper function' constitutes a crucial notion of Millikan's peculiar take on teleosemantics (see Millikan 1984, 1993, 2000). The proper function of a given item, trait or mechanism is what they were *designed for*, what they are *supposed to do* and what they *ought to do*.

There are two aspects of the notion of proper function, which are relevant to our discussion. First, the concept of proper function is a *normative* one, which derives its normativity from its history (Millikan 1984). Thus, the possession by an organism of a given proper function F (in our case, to represent items and facts about the world) is the product of evolution. Second, the cognitive mechanisms producing mental representations are *relational*, that is, they depend on the exchanges and interactions occurring between organism and environment (Millikan 1984, 1993).

On neuroscientific grounds, a teleological theory of content could therefore be defined in the following way. The energetic signals resulting from the organism—environment interactions are transduced and processed in the way they are, in respect of their content, because of the *relevance* (see Sperber 2000) of this content for the possibility of establishing appropriate links between animal behaviour and environment. I would like to emphasize here the 'animal' qualification I purposely assigned to behaviour, in order to make it clear that this approach does not entail and presuppose the necessity of a pre-existing and self-sustaining cognitive framework, whatever it might be and work.

A neuroscientifically grounded teleosemantic approach to conceptual content is, in principle, appealing because it discloses the opportunity to naturalize content in a biologically plausible way. According to this approach, the extension of conceptual content is the way it is not just because of predetermined *a priori* principles (ethereal ideas, God, innate carburettor concepts and the like), but because it is constrained by a general teleological principle. An advantage offered by this approach is that it allows one to describe the evolution of representational content within a naturalistic framework, thus making it empirically tractable.

It can be proposed that the general principle constraining the nature and quality of representational content is probably progressively constructed and diversified by the evolution of more complex patterns of interaction between organisms and their environments. The more these

interactions became articulated, the more they gain complexity by means of their recursive effects on both organisms and their environment.

Given these premises, we now need to characterize the general teleological principle in a more precise and empirically grounded way. My proposal is to equate it to *control strategies*. Any interaction requires a control strategy. Control strategies are typically *relational*: they can be seen as a way of modelling the interaction between organism and environment. However, a model is indeed a form of representation. This step allows a relation of interdependence, if not even superposition, between behaviour control and representation to be established. It also indicates that a moderately counterintuitive domain could be relevant for the naturalization of representational content: the domain of action. I fully address this point in § 4.

Disregarding the problem of what representational content *is made of*, it is indisputable that *some* kind of content, nevertheless, is more useful than other kinds. Properties that are constantly coupled with objects or events, and that reliably occur on different occasions, are most useful, because they enable biological agents to acquire and store knowledge that can also be applied in the future. Furthermore, this kind of 'stored knowledge' allows one to anticipate and predict some properties without the need of always verifying them (see Millikan 2000). Abstraction is exactly that: it enables the representation of objects and events in a way that is independent of their full-blown and constant presence.

Let us now return to the 'nuts and bolts' of neuroscience, and see how representational content can possibly relate to 'localized' brain activity.

3. A NEUROSCIENTIFIC PERSPECTIVE ON THE VEHICLE/CONTENT PROBLEM

Any serious attempt to provide a neuroscientific account of conceptual content, as nested in the activity of the brain, faces the challenge of explaining, among other factors, how the localized patterns of activation of different neural cortical networks can enable the capacity to distinguish, recognize, categorize and, ultimately, conceptualize objects, events and state of affairs in the real world. (I will sidestep possible worlds, and leave them for another occasion.) What can neuroscience tell us about it?

Neuroscience has ultimately taken up this challenge, approaching the problem from its *apparently* easier side, namely the 'neural representation' of feature object concepts. In the interests of brevity, I deliberately do not discuss the important aspect of neurocomputational models, but focus only on the bare experimental empirical evidence.

The recent introduction of brain imaging techniques, such as positron emission tomography, functional magnetic resonance imaging and magnetoencephalography, has provided the opportunity to chart the activation patterns of cortical brain areas with the simultaneous presentations of visual objects as diverse as faces, animals, buildings and places. These results have shown that all these different object categories elicit activity in different, although partly overlapping, regions in the occipito-temporal cortex (see, among others, Martin *et al.* 1996, 2000; Kanwisher *et al.* 1997; Epstein & Kanwisher 1998; Perani *et al.* 1999; Pulvermüller 1999; Kourtzi & Kanwisher 2000, 2001).

A further common characteristic of all these studies is that object categories appear to be represented in more posterior locations within the temporal cortex at a *basic level* (cat, building, face) rather than at a *subordinate level* (the Siamese cat, the Tower of Pisa, Sophia Loren). Conversely, other studies have shown that subordinate level items activate more anterior temporal regions (Damasio *et al.* 1996; Gauthier *et al.* 1997; Gorno-Tempini *et al.* 1998; Leveroni *et al.* 2000).

An important aspect of these brain-imaging studies is that object-related activity has been shown to be consistent across subjects and across different tasks as diverse as object naming, picture matching and word reading. Because a common distinguishing feature of concrete objects is constituted by the shape they have and the way they look, it appeared to be no coincidence that different object categories primarily individuated by their shape (e.g. faces, animals, places) and evoked distinct patterns of activation in those same cortical sectors of the ventral occipito-temporal cortex, which mediate form perception. Thus, this could still only represent a *perceptual* and not conceptual type of generalization.

More recently, single-neuron recording experiments in the human medial temporal cortex of epileptic patients has provided evidence of category-specific visual responses. In a recently published study, over 70% of visually responsive neurons were selectively activated by faces, houses, famous individuals or animals (Kreiman *et al.* 2000). Even more interesting was the result that most of the selective neurons responded only to a single category of stimuli, both during vision and visual imagery. Taken together, this evidence has been used to support the notion that feature object concepts are 'represented' by distributed neural networks that overlap with those involved with object perception (for an excellent and updated review of the relevant literature, see Martin & Chao (2001)).

That said (not much indeed), we are left with the problem of defining *what general principle* exactly constrains the *topology* of the 'representation' of feature object concepts, namely the activity of specific cortical networks. No one can seriously think of a massive one-to-one type of mapping, with distinct brain areas for cats, chairs, cuckoo clocks or palm trees. The problem with this account is that there are too many possible object categories and too little neural space to accommodate all the supposedly discrete and category-specific modules. Several—it should be emphasized—not necessarily mutually exclusive hypotheses have been proposed. Basically, they can be categorized and summarized as follows.

(i) *The category-specific modular hypothesis*. Distinct and specialized cortical modules are supposed to exist, if not for all object categories, at least for *some* of them, such as faces and buildings, supposedly more relevant than others (Kanwisher 2000; Grill-Spector *et al.* 2001).

(ii) *The feature shape hypothesis*. Because objects pertaining to the same conceptual category share many feature characteristics, they tend to be represented

in brain regions sensitive to those same features, organized in a columnar fashion (see Fujita *et al.* 1992).

(iii) *The massively distributed representation hypothesis*. According to this view, the correct interpretation of the extant empirical evidence cannot but suggest that object categories are widely distributed across the visual cortex (Haxby *et al.* 2001).

(iv) *The expertise hypothesis*. The localization constraints do not relate to the intrinsic features of objects, but rather to the expertise of the perceiver. Indeed, it has been shown that novel 'non-face' objects can also activate cortical sites related to face perception, provided the perceiver becomes accustomed to them (Gauthier *et al.* 1999).

As correctly pointed out by Malach *et al.* (2002) in a recent stimulating article reviewing these issues, no one of these hypotheses in itself seems to provide a fully convincing solution to the problems presented. These authors suggest that all the reviewed proposed solutions are insufficient because they are uniquely focused on the *functional* properties of cortical visual areas. An alternative promising strategy—they suggest-could be to consider instead the biased relationship between the retinal eccentricity maps within high—order visual cortices and different categories of objects.

Certain types of objects such as faces, words and letters appear to be associated with a central visual-field bias, while objects such as buildings and places are associated with a peripheral bias. Thus, Malach *et al.* (2002) propose that objects whose recognition and categorization require high visual acuity will be associated with central-biased representations, while objects whose recognition entails large-scale integration will be more peripherally biased.

I think that the solution proposed by Malach *et al.* (2002) is interesting because it relies, though implicitly, on the *relational* aspects of categorization processes. In the account of Malach *et al.* (2002) these relational aspects are uniquely focused on the visual acuity problem. But this is only one side of a multifaceted coin. Let us briefly look at how brain activity correlates with the perception and categorization of *man-made tools*. This will help to fully appreciate the crucial importance of relational factors to shape representational content.

Several brain-imaging experiments have indeed shown that observation, silent naming, and imaging the use of man-made objects leads to the activation of the ventral premotor cortex (Perani *et al.* 1995; Grafton *et al.* 1997; Martin *et al.* 1996; Chao & Martin 2000), a brain cortical region normally considered to be involved in the *control* of action and not in the *representation* of objects. The properties of these objects, that is, their *relational specifications* (how they are supposed to be handled, manipulated and used), appear to comprise a substantial part of their representational content. That explains why the perception of these objects leads to the activation of regions of the brain, which are relevant when we are supposed to interact with those same objects.

The neural *mechanisms* at the basis of the correlation between conceptual knowledge and premotor activation should now be clarified. A very useful strategy is to rely on the level of description with high spatial and temporal resolution, which the direct correlation between the activity of *single neurons* and the parallel ongoing behaviour of the organism provides. This is the level of description at which cognitive neurophysiology operates.

4. THE NEUROBIOLOGY OF GOALS

Neurophysiology has been, for decades, patently reluctant to engage itself in any research programme promoted to delve into the realm of the intentional/representational aspects of behaviour. The target of neurophysiological research entailed the investigation of sensory processes such as vision and output functions, for example motor behaviour. Motor behaviour, in turn, was, and by some researchers still is, uniquely envisaged as a multilayered process to be studied and characterized *exclusively* in terms of very elementary physical features such as force, direction and amplitude. Even without any explicit commitment to investigate the possible cognitive entailments of the neural control of motor behaviour, a set of empirical results nevertheless almost forces us to cope with the previously neglected cognitive aspects of action and its control.

I will illustrate a series of empirical evidence that points, forcefully, to a crucial role played by *interaction* in shaping, defining and constraining the representational aspects of the dynamic interplay between organisms and environment. To achieve this, I introduce the neural properties of a sector of the premotor cortex of macaque monkeys studied in our laboratory for more than 20 years.

The rostral-most sector of the ventral premotor cortex of the macaque monkey controls hand and mouth movements (Rizzolatti *et al.* 1981, 1988; Kurata & Tanji 1986; Hepp-Reymond *et al.* 1994). This sector, which has specific histochemical and cytoarchitectonic features, has been termed area F5 (Matelli *et al.* 1985). A fundamental functional property of area F5 is that most of its neurons do not discharge in association with elementary movements, but are active during *actions* such as grasping, tearing, holding or manipulating objects (Rizzolatti *et al.* 1988).

What is coded here is not simply a physical parameter of movement such as force or movement direction, but rather the relationship, in motor terms, between the agent and the object of the action. Furthermore, this relation is of a very special kind: a relation leading to success. A hand reaches for an object, grasps it, and does things with it. F5 neurons indeed become active only *if* a particular type of agentobject relation (e.g. hand–object) is executed until the relation leads to a different state (e.g. to take possession of a piece of food, to throw away an object, to break it, to bring it to the mouth). Particularly interesting in this respect are grasping-related neurons that fire whenever the monkey *successfully* grasps an object, regardless of the effector employed, be it either of his two hands, the mouth or both (Rizzolatti *et al.* 1988; see also Rizzolatti *et al.* 2000).

The independence between the nature of the effector involved and the end-state that the same effector is supposed to attain constitutes an *abstract* kind of means–end representation. We can envisage it as the dawning of more sophisticated articulations to come.

The presence in the motor system of a specific neural

format for states that could be representationally defined as *concepts of action goals* allows for a much simpler selection of a particular action within a given context (Rizzolatti *et al.* 1988). Both when the action is self-generated and when it is externally driven, only a few representational elements need to be selected.

Within the context of a *motor, interactive* representational code for goals, motor acts aimed at a specific goal can be represented in the brain just as such, as goal states, and not in the far less economical terms of the specification and control of individual movements. Thus, to categorize a teleological concept, we have a representational neural format that generalizes across different instances in which a particular successful end-state of the organism (the goal) can be achieved. In accord with information theory, the conceptual narrower state has been reached by getting rid of useless, redundant information (for example, the load of information about *all* the dynamic patterns under which an intentional action can be characterized).

Beyond purely motor neurons, which constitute the overall majority of all F5 neurons, area F5 also contains two categories of 'visuomotor' neurons. Neurons of both categories have motor properties that are indistinguishable from those of the above-described purely motor neurons, while they have peculiar 'visual' properties. The first category comprises neurons responding to the presentation of objects of a particular size and shape in the absence of any detectable action aimed at them, either by the monkey or by the experimenter. These neurons have been defined as 'canonical neurons' (Rizzolatti & Fadiga 1998; Rizzolatti *et al.* 2000).

The second category is made by neurons that discharge when the monkey *observes* an action made by another individual and when it *executes* the same or a similar action. We called these latter visuomotor neurons 'mirror neurons' (Gallese *et al.* 1996; Rizzolatti *et al.* 1996; for a recent review, see Rizzolatti *et al.* 2001).

I will confine myself to the discussion of 'canonical neurons'. Let us have a closer look at them. Because most grasping actions are executed under visual guidance, a relationship has to be established between the features of 3D visual objects and the specific motor specifications they might engender *if* the animal is aiming at them. The appearance of a graspable object in the visual space will retrieve immediately the appropriate 'motor representation' of the intended type of hand–object relation. This process, in neurophysiological terms, implies that the same neuron must be able not only to code the motor acts it is supposed to control, but also to respond to the situated visual features triggering them.

Indeed, 'canonical neurons' respond to the visual presentation of objects of different size and shape in the absence of any detectable movement of the monkey (Rizzolatti *et al.* 1988, 2000; Jeannerod *et al.* 1995; Murata *et al.* 1997). Frequently, 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, however, is the fact that in a considerable percentage of neurons the congruence is observed between the specification of a given type of 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 first conclusion is that such neurons contribute to a multimodal representation of an organism–object relation.

The function of F5 canonical grasping neurons can therefore hardly be defined in purely sensory or motor terms. At this stage, object representations seem to be processed in 'relationally specified' terms (Gallese 2000*a*,*b*). Within the operational logic of such neural network, a series of physical entities, 3D objects, are identified, differentiated and conceptualized, not in relation to their mere physical appearance, but in relation to the effect of the interaction with an acting agent. I regard this as a good example of an intentional type of representation: specifically and exclusively coded under a distinct type of neural activity patterns, one involving dynamic organism–object relations.

The experiments reviewed above shed light on the neural mechanism at the basis of the unexpected correlation disclosed in humans by brain-imaging techniques between categorical perception of tools and the activation of premotor brain sectors, thus giving further empirical grounding to my previous proposal.

I should emphasize, however, that I am not willing to say that the sole constituent of the representational content of a given object is its *manipulability* value. What the limited 'vocabulary' of actions (Rizzolatti *et al.* 1988; see also Rizzolatti *et al.* 2000) represented in area F5 suggests, nevertheless, is that the intentional character, the 'aboutness' of the representational format of our mind, is deeply rooted in the intrinsic relational character of body action, which, in turn, suggests the intrinsic intertwined character of action, perception and cognition (Gallese 2000*b*).

Representational content, and thus—*a fortiori*—conceptual content, cannot be fully explained without considering it as the result of the ongoing modelling process of an organism as currently integrated with the object to be represented, by intending it. This integration process between the representing organism and the represented object is articulated in a multiple fashion, for example, by intending to explore it by moving the eyes, intending to hold it in the focus of attention, by intending to grasp it, and ultimately, by *thinking* about it (see Gallese 2000*b*; Gallese & Metzinger 2003; see also Metzinger 1993, 2000, 2002).

I think that the proposed equivalence between control strategies and representation should now be clearer. The intrinsic need of any organism to control its dynamic interaction with the environment also constrains the way these interactions are supposed to be modelled and henceforth represented. Nature seems to have operated during the course of evolution according to a principle of parsimony. The same *sensorimotor* circuits that control the ongoing activity of the organism within its environment also map objects and events in that very same environment, thus defining and shaping their representational content. It is no coincidence that our representation of the world is a *model* of it that *must* incorporate our idiosyncratic way to interact with it. This stems from the peculiar and unique way biological organisms are supposed to gain information about the world, that is, by transducing its energetic nature into neural action potentials, through a peculiar type of active interaction with the world, in turn constrained by how living organisms' bodies are built and

how the world is. Our perspective on the reality of the world cannot be but a *model of the world* for that very reason.

It should now also be clearer why vehicle and content of representation should not be qualified as ontologically independent entities. Mammals, because of the way they are, can only represent the world by modelling it. We have learned also that this model can only be conceived as an integrated dynamic interplay between situated organisms and their natural playground. It follows, from that, that the representational content resulting from the use of neural information for control purposes, and that same neural information, both share the same ontological status. It must be emphasized, however, that such equivalence only holds if we qualify neural information as constrained and determined by the peculiar nature of the organisms making use of it. Simply, the producer and the repository of representational content is not the brain *per se*, but the entire organism, by means of its interactions with the world of which it is a part.

5. DO MONKEYS HAVE CONCEPTS?

As initially stated, the current dominant view in philosophy of mind and cognitive science precludes to non-human animals the possession of concepts and abstraction. One of the most frequently used arguments to substantiate this preclusion is lack of *consciousness of content*. This argument can be defined in the following way: for a system S to have a concept Y of X requires S to be aware of the causal relations between Y and X, and of the semantic content of Y (see Stich 1978; Searle 1980).

Most of what is going on 'inside' our cognitive system, however, is *transparent* to our knowledge of it. This means that we are constantly using inferential strategies and applying conceptual knowledge without the need to be aware of what occurs 'inside our head', either in terms of scientific knowledge about how the brain works, or in terms of the constant application of self-directed metacognition. It is extremely unlikely that to entertain the concepts of 'chair' or 'goal' we explicitly need to tag them as concepts. It is also easy to see why things work that way. Imagine what a nightmarish experience it would be if the thermostat of your heating system shouted out, every second, the temperature it was measuring. We have conceived and built control systems in order not to bother doing the things they do for us. In the course of the evolution of biological cognitive systems, nature might well have operated by using the same logic.

Furthermore, there is perhaps a deeper reason to suggest that we should conceive the capacity to have concepts as being completely independent from the capacity to explicitly represent the *content* of a concept *as a concept*. The reason is that infinite regress is *biologically* impossible *and* logically unsound. All poorly epistemically equipped theories about the mind, when defining thought in terms of mental representations, face the risk of infinite regress. An internal conceptual representation of a thing in the world as entertained by a biological organism cannot be supposed to do its semantic job, and simultaneously internally represent its semantic, at pain of infinite regress. The transparency of content *as* content is the rule not the exception. It is only when an incredibly more powerful 'Mode of Presentation' of semantic content (see Fodor 1998) becomes available, when language finally appears in evolution, that conceptual content can also be explicitly represented as such.

The investigation of the epistemic relation between organism and environment is the only way to define the causal and nomological relation between mental representations and objects and facts in the world. The functional architecture that neuroscientific empirical research is beginning to unravel is that of a control strategy based on multiple specifications of relations. Language, hence the capacity to make conceptual content *opaque*, is just the expression of a more sophisticated Mode of Presentation whose identity is constituted by '...what happens when you entertain it' (Fodor 1998, p. 20).

From this logic, it must follow that the absence of self-referential sophisticated cognitive capacities does not necessarily preclude the organisms devoid of them from entertaining conceptual knowledge and from using it in guiding their own behaviour. This means that in the cognitive domain, what makes humans and animals different is the level of complexity of their functional control-logic. The 'no-preclusion' condition does not presuppose, however, that non-human animals do indeed have and make use of concepts. Eliminating the preclusion enables these issues to be empirically treated. Hence, animal concepts can legitimately become the target of scientific inquiry.

Let us have a closer look at language in relation to the now empirically tractable problem of whether non-linguistic species do have concepts. It is easy to understand why it is intuitively plausible to consider language paramount for the possession of concepts. Language is so powerful a cognitive tool simply because it enables *sameness of content* in spite of the potential multiplicity of states subsuming it. It is by far the most powerful 'generalization device' our cognitive system is equipped with.

Interestingly enough, sameness of content as resulting from a multiplicity of states subsuming it is, however, also a necessary condition for concepts, in order to guarantee them the capacity to misrepresent. As shown by Dretske (1986), it is only when information is conveyed through multiple paths connecting a fact in the world F with an internal representation R that R can also misrepresent F.

Should we, then, grant the capacity to have 'conceptual knowledge' about objects and facts in the world to non-linguistic species such as monkeys, if we could show that their brain can represent a specific object or fact in the world through multiple AND converging sensorimotor paths? I think that many would be ready to answer yes. The evidence on canonical neurons presented above suggests that this might be the case. However, some recent neurophysiologycal data obtained in our laboratory seem to really make the case.

As demonstrated in § 4, in the monkey premotor cortex (area F5) there are neurons, mirror neurons, that discharge both when the monkey makes a specific action and when it observes another individual making a similar action (Gallese *et al.* 1996; Rizzolatti *et al.* 1996). In a recent study, we investigated whether there are neurons in F5 that discharge when the monkey makes a specific hand action and also when it *hears* the corresponding action-related sounds. The results show that monkey premotor cortex is equipped with neurons that discharge

when the monkey *executes* an action, *sees* or just *hears* the same action performed by another agent (see Kohler *et al.* 2001, 2002). We called these neurons 'audio-visual mirror neurons' (Kohler *et al.* 2002). They respond to the sound of actions and discriminate between the sounds of different actions, but do not bother to respond to other similarly interesting sounds such as arousing noises, or monkeys' and other animals' vocalizations. The actions whose sounds are preferred are also the actions producing the strongest responses when observed or executed. It does not differ significantly for the activity of this neural network if matters of fact of the world such as a peanut being broken, or a sheet of paper being torn apart, are specified at the motor, visual or auditory level. Such a neural mechanism enables representation of the consequences of an action, thus its goal, in a way that is in principle also open to misrepresentation (e.g. neurons responding to a sound different from that produced by the action coded by them when executed or observed). Furthermore, the same conceptual content ('the goal of action A') results from a multiplicity of states subsuming it, namely, differently triggered patterns of activation within a population of 'audio-visual mirror neurons'.

If, as I am proposing, 'audio-visual mirror neurons' instantiate a conceptualization mechanism, it is open to future empirical investigation whether such a mechanism can acquire further levels of generalization by means of the learning process.

6. CONCLUSIONS

The results briefly reviewed above concerning 'audio-visual mirror neurons' are important in that they seem to suggest that it is possible to have sameness of informational content at a quite 'abstract' level, the level of conceptual content, without being endowed with the cognitive faculty of language. If the different mode of presentation of events as intrinsically different as sounds, images, or willed effortful acts of the body is nevertheless bound together within a circumscribed, informational lighter level of semantic reference, what we have here is a *mechanism instantiating conceptualization.* Abstraction appears therefore to be possible also for organisms that, as monkeys, are devoid of linguistic abilities.

This reopens, perhaps from an unusual perspective, the problem of the relation between thought and language. The concepts that the system computes are neither 'linguistic' nor 'symbolic'. The level of conceptual knowledge that I am proposing to ascribe to monkeys, (but that I also take to be still quite alive in our human mind), thus enabling them with the possibility to entertain abstract contents, is heavily dependent on 'implicit inferences'. Inferences are just more or less reliable predictions about facts. And prediction is a product of the constant reshaping/rewiring of our models of the world as we interact with it. Thus, the implicit inferences produced by internal models are the driving force that builds concepts, thus enabling abstraction.

Concepts just contribute to make a certain kind of thought possible. It should also be noted that because the capacity to predict facts and events is one of the leading capacities that make a cognitive system really smart, the involvement of model-driven implicit inference in the determination of conceptual content in no way diminishes and undermines its high cognitive status.

The investigation of the neural mechanisms at the basis of the relationally specified transactions between organism and world appears to be a very promising source of information for the difficult task of naturalizing concepts and understanding what underpins our capacity for abstraction.

Shorter and preliminary versions of this paper have been presented at the Munich Encounters in Cognition and Action at the Max Planck Institute in Munich, Germany, in December 2000; at the 11th International Meeting on the Neural Control of Movement held in Sevilla, Spain, in April 2001; at the international workshop of the ACI Cognitique, 'L'Abstraction Dans La Cognition Animale', held in Paris, France, in April 2001; at the Second Meeting of the McDonnell Project in Philosophy and the Neurosciences, held in Tofino, Canada, in June 2001; and at the International Workshop on 'Concepts' organized by the Department of Communication Sciences of the University of Bologna, Italy, in May 2002. The author thanks all audiences for the feedback received from them. Particular thanks to George Lakoff, Jerry Feldman, Giacomo Rizzolatti and Thomas Metzinger for their invaluable help in discussing some of the issues addressed in this paper, and to all members of the McDonnell Project in Philosophy and the Neurosciences directed by Kathleen Akins, for the stimulating experience of interdisciplinary discussions. Supported by the Eurocores Programme of the European Science Foundation and by MIURST.

REFERENCES

Brentano, F. 1973 (1874) *Psychologie vom empirischen Standpunkt*. Erster Band. Hamburg: Meiner. [English translation: Brentano, F. 1874 (1973). *Psychology from an empirical standpoint* (ed. O. Kraus). English edition edited by Linda McAlister, translated by A. C. Rancurello, D. B. Terrell and L. McAlister. London: Routledge & Kegan Paul/New York: Humanities Press].

Chao, L. L. & Martin, A. 2000 Representation of manipulable man-made objects in the dorsal stream. *Neuroimage* **12**, 478–484.

Chomsky, N. 1986 *Knowledge of language: its nature, origin and use*. New York: Praeger Special Studies.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D. & Damasio, A. R. 1996 A neural basis for lexical retrieval. *Nature* **380**, 499–505.

Dretske, F. 1981 *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. I. 1986 Misrepresentation. In *Belief, form, content, and function* (ed. R. J. Bogdan), pp. 17–34. Oxford: Clarendon.

Dretske, F. 1988 *Explaining behavior*. Cambridge, MA: Bradford Books/MIT Press.

Dretske, F. 1995 *Naturalizing the mind*. Cambridge, MA: MIT Press.

Epstein, R. & Kanwisher, N. 1998 A cortical representation of the local visual environment. *Nature* **392**, 598–601.

Evans, G. 1982 *The varieties of reference*. Oxford: Clarendon.

Fodor, J. 1975 *The language of thought*. New York: Thomas Y. Crowell.

Fodor, J. 1981 *Representations*. Cambridge, MA: MIT Press.

Fodor, J. 1987 *Psychosemantics: the problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. 1998 *Concepts: where cognitive science went wrong*. Oxford Cognitive Science Series. Oxford: Clarendon.

Fodor, J. 2001 *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Fujita, I., Tanaka, K., Ito, M. & Cheng, K. 1992 Columns for visual features of objects in monkey inferotemporal cortex. *Nature* **360**, 343–346.

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

Gallese, V. 2000*b* The inner sense of action: agency and motor representations. *J. Consc. Stud.* **7**, 23–40.

Gallese, V. & Metzinger, T. 2003 Motor ontology: the representational reality of goals, actions and selves. *Phil. Psychol.* (In the press.)

Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. 1996 Action recognition in the premotor cortex. *Brain* **119**, 593–609.

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P. & Gore, J. C. 1997 Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Curr. Biol.* **7**, 645–651.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P. & Gore, J. C. 1999 Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neurosci.* **2**, 568–573.

Grafton, S. T., Fadiga, L., Arbib, M. A. & Rizzolatti, G. 1997 Premotor cortex activation during observation and naming of familiar tools. *Neuroimage* **6**, 231–236.

Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. 2001 The lateral occipital complex and its role in object recognition. *Vision Res.* **41**, 1409–1422.

Gorno-Tempini, M. L., Price, C. J., Josephs, O., Vandeberghe, R., Cappa, S. F., Kapur, N. & Frackowiak, R. S. J. 1998 The neural systems sustaining face and proper name processing. *Brain* **121**, 2103–2118.

Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. 2001 Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430.

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. *Can. J. Physiol. Pharmacol.* **72**, 571–579.

Jeannerod, M., Arbib, M. A., Rizzolatti, G. & Sakata, H. 1995 Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci.* **18**, 314–320.

Kanwisher, N. 2000 Domain specificity in face perception. *Nature Neurosci.* **8**, 759–763.

Kanwisher, N., McDermott, J. & Chun, M. M. 1997 The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311.

Kohler, E., Umiltà, M. A., Keysers, C., Gallese, V., Fogassi, L. & Rizzolatti, G. 2001 Auditory mirror neurons in the ventral premotor cortex of the monkey. *Soc. Neurosci. Abstr.* **XXVII**, 129.9.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V. & Rizzolatti, G. 2002 Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**, 846–848.

Kourtzi, Z. & Kanwisher, N. 2000 Cortical regions involved in perceiving object shape. *J. Neurosci.* **20**, 3310–3318.

Kourtzi, Z. & Kanwisher, N. 2001 Representation of perceived object shape by the human lateral occipital complex. *Science* **293**, 1506–1509.

Kreiman, G., Koch, C. & Fried, I. 2000 Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neurosci.* **3**, 946–953.

Kurata, K. & Tanji, J. 1986 Premotor cortex neurons in macaques: activity before distal and proximal forelimb movements. *J. Neurosci.* **6**, 403–411.

Lakoff, G. 1987 *Women, fire, and dangerous things*. Chicago University Press.

Lakoff, G. & Johnson, M. 1980 *Metaphors we live by*. Chicago University Press.

Lakoff, G. & Johnson, M. 1999 *Philosophy in the flesh*. New York: Basic Books.

Lakoff, G. & Nuñez, R. 2000 *Where mathematics comes from: how the embodied mind brings mathematics into being*. New York: Basic Books.

Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R. & Rao, S. M. 2000 Neural systems underlying the recognition of familiar and newly learned faces. *J. Neurosci.* **20**, 878–886.

Malach, R., Levy, I. & Hasson, U. 2002 The topography of high-order human object areas. *Trends Cogn. Sci.* **6**, 176–184.

Martin, A. & Chao, L. L. 2001 Semantic memory and the brain: structure and processes. *Curr. Opin. Neurobiol.* **11**, 194–201.

Martin, A., Wiggs, C. L., Ungerleider, L. G. & Haxby, J. V. 1996 Neural correlates of category-specific knowledge. *Nature* **379**, 649–652.

Martin, A., Ungerleider, L. G. & Haxby, J. V. 2000 Category-specificity and the brain: the sensory-motor model of semantic representations of objects. In *The new cognitive neurosciences*, 2nd edn (ed. M. S. Gazzaniga), pp. 1023–1036. Cambridge, MA: MIT Press.

Matelli, M., Luppino, G. & Rizzolatti, G. 1985 Patterns of cytochrome oxidase activity in the frontal agranular cortex of the macaque monkey. *Behav. Brain Res.* **18**, 125–137.

Metzinger, T. 1993 *Subjekt und Selbstmodell*. Padeborn: Schoeningh.

Metzinger, T. 2000 The subjectivity of subjective experience: a representationalist analysis of the first-person perspective. In *Neural correlates of consciousness. Empirical and conceptual questions* (ed. T. Metzinger), pp. 285–306. Cambridge, MA: MIT Press.

Metzinger, T. 2002 *Being no one. The self-model theory of subjectivity*. Boston, MA: MIT Press.

Millikan, R. G. 1984 *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Millikan, R. G. 1993 *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

Millikan, R. G. 2000 *On clear and confused ideas. An essay on substance concepts*. Cambridge University 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. *J. Neurophysiol.* **78**, 2226–2230.

Papineau, D. 1987 *Reality and representation*. Oxford: Blackwell.

Perani, D., Cappa, S. F., Bettinardi, V., Bressi, S., Gorno-Tempini, M., Matarrese, M. & Fazio, F. 1995 Different neural systems for the recognition of animals and man-made tools. *Neuroreport* **6**, 1637–1641.

Perani, D., Schnur, T., Tettamanti, M., Gorno-Tempini, M., Cappa, S. F. & Fazio, F. 1999 Word and picture matching: a PET study of semantic category effects. *Neuropsychologia* **37**, 293–306.

Pulvermüller, F. 1999 Words in the brain's language. *Behav. Brain Sci.* **22**, 253–279.

Pylyshyn, Z. W. 1984 *Computation and cognition: toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Rizzolatti, G. & Fadiga, L. 1998 Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). In *Sensory guidance of movement. Novartis Foundation Symp.* 218, pp. 81–103. Chichester: Wiley.

Rizzolatti, G., Scandolara, C., Gentilucci, M. & Camarda, R. 1981 Response properties and behavioral modulation of 'mouth' neurons of the postarcuate cortex (area 6) in macaque monkeys. *Brain Res.* **255**, 421–424.

Rizzolatti, G., Camarda, R., Fogassi, M., 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. *Exp. Brain Res.* **71**, 491–507.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. 1996 Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* **3**, 131–141.

Rizzolatti, G., Fogassi, L. & Gallese, V. 2000 Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In *The cognitive neurosciences*, 2nd edn (ed. M. S. Gazzaniga), pp. 539–552. Cambridge, MA: MIT Press.

Rizzolatti, G., Fogassi, L. & Gallese, V. 2001 Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Neurosci. Rev.* **2**, 661–670.

Searle, J. R. 1980 Minds, brains, and programs. *Behavioural Brain Sci.* **3**, 417–424.

Sperber, D. 2000 Metarepresentations in an evolutionary perspective. In *Metarepresentations: a multidisciplinary perspective. Vancouver Studies in Cognitive Science*, vol. 10 (ed. D. Sperber), pp. 117–138. Oxford University Press.

Stampe, D. 1977 Toward a causal theory of linguistic representation. In *Midwest studies in philosophy: studies in the philosophy of language*, vol. 2 (ed. P. A. French, T. E. Uehling Jr & H. K. Wettstein), pp. 81–102. Minneapolis, MN: University of Minnesota Press.

Stich, S. 1978 Beliefs and suboxastic states. *Philosophy Sci.* **45**, 499–518.