HUMANA MENTE

ISSUE 15 - JANUARY 2011

Agency:

FROM EMBODIED COGNITION TO FREE WILL

EDITED BY DUCCIO MANETTI AND SILVANO ZIPOLI CAIANI

Edizioni ETS

Editorial Board

Editorial Staff

Table of Contents

BOOK REVIEWS

COMMENTARIES

Introduction
Agency: From Embodied Cognition to Free Will

Duccio Manetti *
duccio.manetti@unifi.it

Silvano Zipoli Caiani **
silvano.zipoli@unimi.it

Traditional theories about experience have always represented the subject as a *passive* recipient of sensory stimuli, which get processed through successive layers of the brain cortex and culminate in a phenomenal experience, omitting any mention of the role of the personal sense of agency. According to this formulation, experience emerges as a combination of biological and phenomenological descriptions, linking mechanical processes to subjective qualitative reports. Conceptual frameworks provided by neuroscience and phenomenological analysis are alternative descriptive systems originally conceived for alternative explanatory purposes. Here is the origin of many of the theoretical tensions in cognitive science. Today, after years in which *dualism* and *reductionism* have been the only games in town, the idea of an embodied dynamicism is emerging in the field of cognitive science with support from substantial empirical evidence. As perceptual experience is shaped by action execution, it seems necessary to assume a theoretical framework within which the interconnection between the perceiving subject's *conscious states*, his *body* and the *environment* is adequately emphasized.

For the phenomenological debate, the notion of embodiment coincides with the rebuttal of what is usually considered the *Cartesian dualism*, that is, the segregation of any bodily influence from the subjective experiential domain. Crossing the history of western thought, this problem acquires a critical dimension in the twentieth century philosophical debate. The way to understand the relationship between body and consciousness finds a new style after the establishment of the phenomenological framework. Following the path originally drawn by Husserl and successively developed by Merleau-Ponty,

* University of Florence
** University of Milan

it is possible to figure out how the phenomenological tradition, from its early stages, has originally approached the mind-body problem underlying the opportunity to develop an *interactive* conception based on the assumption of a radical *interweaving* between the *experiential* and the *bodily* domains.

According to this view, perceptive experience can be conceived as a method through which the subject travels in the environment following his motor intentions and exploiting his skillful knowledge of the sensorimotor constraints that link the execution of a goal oriented action to the variation of the phenomenal features.

Working on the clarification of the notion of embodiment we have the opportunity to cease to unreflectively privilege only one possible explanation of our experience. The human mind, observed through the lenses of embodiment, emerges at the interface of the brain, the body, the material and social environment. This is an inextricable mash influencing all aspects of our life. We are agents whose nature is fixed by a complex interaction involving our personal experience, a particular kind of physical embodiment and a certain embedding in the environment. This very combination of *experience*, *flesh* and *environment* is the main character of our being in the world.

The assumption of agency as a critical aspect of our experience motivates the introduction of another classical philosophical problem such as that concerning the notion of free will. We usually consider human beings natural organisms that are morally responsible for their own actions. Yet this assumption represents one of the most intriguing puzzles that, from ancient Greece to the contemporary era, has absorbed philosophers and scientists of every kind. Are we really free agents? What does our subjective experience of agency reveal to us? And how do these questions relate to the fact that we are natural embodied beings?

Except in cases where we are physically constrained, we consider ourselves free beings that think, believe and act autonomously, that is, according to the states of consciousness that characterize our own mental life. We consider ourselves responsible for our own acts because we perceive ourselves as being able to freely project the actions that our body can perform. Accordingly, the possibility of a free choice appears to be strictly related to the possibility of assigning independence to a particular domain such as our subjective consciousness.

The subjective sense of agency, that is, the feeling that we control our own movements and actions, is certainly an essential, constant element of our

everyday experience. It seems obvious to us that the casual chain leading to the execution of an action critically derives from our conscious intention. However, we can try for a moment to imagine we do not have any real power over our actions. We can imagine that we are prisoners of an illusion that gives us the impression that we are the causes of our actions, but that we are actually nothing but automata governed by a sophisticated system of behavioral laws. If we carried through with this imaginative effort, then the very meaning of the word "freedom" would need to be modified according to the idea that those we perceive as our voluntary actions are, in reality, independent of our will. But does this make sense? Or is it only a philosophical trick?

The aim of the present issue of Humana.Mente is to frame the debate by introducing original arguments in the fields of theory of agency and free will. With this purpose in mind, we invited authors from different disciplines to submit their contributions. We received enthusiastic replies from some of the most prominent scholars working in these fields. This is certainly evidence that the topic we proposed still arouses steady interest even after over two thousand years of philosophical and scientific discussion. This volume is also evidence that the debate is not frozen and that new conceptions and perspectives have been developed over the last ten years. In order to make the composition of the issue clear, we decided to divide the Papers Section into two parts. The former devoted to introduce arguments concerning the theory of agency, the latter devoted to introduce specific perspectives on the notion of free will. Now, let us briefly illustrate the content of the volume.

The opening paper by Michael Silberstein and Antony Chemero is an introduction to a dynamical account of intentional actions and agency. Silberstein and Chemero contrast the idea that action is caused by disembodied mental representations residing in the head and move from the assumption that cognitive systems are genuinely *extended* structures, which effectively connect the *brain* to the *body* and to the *environment*. Following this line of thought, the body and the environment can be considered a continuous *dynamical system* constituted by variables that change according to mathematical laws. This makes it possible to account for cognitive processes through differential equations that pair animal parameters with environmental parameters. It is important to note that, in light of its radical *anti-representationalism* and *anti-computationalism,* Silberstein and Chemero's dynamical theory constitutes a special approach to the *extended mind*

paradigm, different from other proposals in this field (e.g., Clark's conception). Indeed, the proposal advocated by the authors is in continuity with Gibson's ecological psychology, according to which cognition and conscious experience are ongoing adaptive activities performed by animals in their natural niche. According to this view, actions and environmental conditions influence each other, such that the agent and the environment can be viewed as two co-dependent sides of the same coin.

As a kind of enactive approach to agency, Torrance and Froese's paper also focuses on the dynamics of agents interacting with the environment. More precisely, the environment is characterized as a system of conditions and constraints imposed by a social situation where agents interact with each other. Accordingly, the authors argue against what they call "methodological solipsism" in cognitive science, emphasizing the role of historical and social norms in shaping our subjective experience of agency. The authors discuss many examples from common experience and artificial intelligence, showing how the (relative) autonomy of an interaction process, which is separate from the autonomy of individual participants, has the power to influence an agent's individual goals. Accordingly, the main challenge of the paper is to show how social interactions actually co-constitute the individual's sense of agency, as well as how the individual's actions are involved in the constitution of social situations.

The role of social interaction in the formation of a sense of agency is also emphasized by Shaun Gallagher. Gallagher's paper criticizes the standard debate in theory of mind, which is characterized by a dispute between theory-theory and simulation theory; Gallagher defends an alternative approach that he calls *interaction theory*. According to Gallagher, interaction theory faces many suppositions associated with the traditional approach in theory of mind, arguing for three basic assumptions. First of all, other minds are not hidden, inaccessible entities, but become manifest through other people's behavior. Second, Gallagher assumes that our everyday stance toward other people is not merely a detached observation; rather, it is almost always the result of embodied interactions and communicative actions. Finally, in Gallagher's view, understanding others doesn't involve a process of mentalizing; it is a direct and spontaneous activity that characterizes our life. In this paper, Gallagher introduces a developmental model according to which adult communicative and narrative practices – such as sensory-motor abilities (primary intersubjectivity), joint attention and pragmatic engagement (secondary

intersubjectivity) – develop from strong embodied interaction with other people. According to Gallagher, autonomy is not an "internal and intra-individual negotiation", but it is the expression of the way people arrange their lives with others. Following this line, self agency emerges as a characteristic defined by the network of human relationships, instead of a purely individual attribute.

Next, Horgan's paper argues about the phenomenology of agency and its consequences on the freedom-determinism debate. In the first section of the paper, the author introduces some features of agentive phenomenology as made available by introspective attention. Horgan's analysis is particularly concerned with what he considers the erroneous presupposition that any genuine phenomenal question can be reliably answered directly through introspection, tempting one to think that introspection alone can solve every dilemma concerning the nature of the subjective experience of agency. On the contrary, Horgan argues, the self is inadequate as an ultimate source to find the answers to questions about the nature of agency and freedom. Accordingly, using an *abductive* argument, Horgan attempts to show why we cannot reliably ascertain the nature of agency based solely on careful introspection, due to our strong natural tendency to judge freedom as an essential and evident component of our experience of acting.

Our subjective experience of agency, like various cognitive processes, is shaped by specific *bodily constraints*. The way in which an organism is embodied determines how a subject interacts with specific aspects of the environment, thus influencing the rise of sensory-motor experiences which serve as the basis for the formation of categories and concepts concerning our phenomenology of action.

Accordingly, Mauro Maldonato highlights the unconscious role of the body in agency dynamics. In the author's opinion, even if we are normally led to emphasize the role of perception and sensation, assuming that our voluntary movements are essentially dependent on them, our phenomenology of action is rooted in the motor system itself. Maldonato's analysis focuses on the negative consequences derived from the traditional separation of mental functions from bodily dimension, drawing from many examples in the field of neurobiology to show how the mind is profoundly influenced by the motor sphere. According to Maldonato, motility has not only direct and overt consequences, but also critical effects on other cognitive systems, such as those underlying perception and language understanding. This conception shows that the boundary

between action and perception is not as sharp as it is usually supposed to be, and that a great deal of cognition can be surprisingly related to the functioning of the agent's motor system.

Phenomenology of agency cannot be divorced from the critical question of how we can actually control our voluntary behavior, or from the question concerning the existence of a causal link between our feeling about performing a specific action and the action itself. Accordingly, the second part of the Papers Section includes contributions that introduce new aspects and perspectives concerning the *vexata quaestio* of free will. Today, now that refined techniques of enquiry in the field of neuroscience have been developed, participants in the free will debate are particularly engaged in interpreting the increasing amount of empirical data, which seems to threaten the traditional dichotomy between deterministsand libertarians. An example of this tendency is visible in the interest that Libet's experiments still arouse in both the scientific and the philosophical communities. Over the years Libet's experimental paradigm has become a critical topic where the interests of contrasting positions converge.

Given this trend, we decided to encourage contributions on free will concerning the interpretation of empirical findings and the development of theoretical frameworks. In keeping with this intention, for this section we collected papers from prominent scholars in philosophy, psychology and neuroscience. The overall result gives the reader a taste of how many different approaches and styles characterize this fascinating debate. The first paper, by Roberta De Monticelli, begins with an introduction to phenomenology as the method based on "epistemic trust" in the world of experience, having the power to characterize things as irreducible to their psychological, biological and physical constitution. According to the author, the question of free will can be considered as a genuine matter of epistemic trust, that is, of reliability concerning *ordinary* experience. De Monticelli's point is that, in order to become a subject of acts and develop selfhood, one must entertain a relationship of epistemic adequacy with the phenomenal world. Accordingly, distinguishing between two orders of positionality, the author shows how the persistence of the problem of free will depends on a sort of fallacy in the order of explanation.

The paper by Davide Rigoni, Luca Sammicheli and Marcel Brass critically discusses a series of influential experiments in the field of cognitive neuroscience, concerning the relationship between the subjective sense of

agency and the actual execution of intentional actions. The authors' analysis refers to a large amount of data according to which the execution of motor actions is always preceded by unconscious brain processes; the individual's subjective experience of conscious intentions is purportedly inferred from the event occurring after the action is executed. Results of this kind challenge the intuitive view that we are responsible for the actions we execute, as our conscious intention to act appears to be an unessential component. Notwithstanding this empirical evidence, the authors' point is that considering free will as a mere epiphenomenal illusion would be an overstatement. To support this claim, Rigoni, Sammicheli and Brass focus on our natural tendency to perceive free will in others, emphasizing the underestimated pragmatic value of believing in freedom rather than in determinism.

Susan Pockett's paper frames the free will debate by introducing some implications related to the assumption of what she calls *electromagnetic field theory of consciousness*. This is an identity theory according to which consciousness is identical to specific electromagnetic field patterns induced by neural activity. Unlike other materialist identity theories, Susan Pockett's theory doesn't assume a causal link between the electromagnetic fields and the initiation of bodily movements. On the contrary, Pockett defends an electromagnetic field theory of consciousness citing crucial reasons for rejecting the belief that consciousness causes bodily movements and, therefore, for rejecting the claim that electromagnetic patterns are involved in our subjective experience of agency.

In the next paper, Bickhard proposes a radical critique of a computational model of decision-making, where actions are the final elements of a causal chain made of many point-like events through which the causal influence is transmitted. According to this view, a decision to act is a computational process that starts with a reason and ends with a motor execution. In contrast to this view, Bickhard assumes that decision and action are two aspects of the same underlying kind of process. Rejecting a pointillist picture, the author defines a decision to act as a temporally extended and self-organizing process. According to this view, Bickhard's model of acting is determined by global characteristics instead of reducible local causal attributes.

Jing Zhu's paper supports a libertarian approach to the question of free will according to which indeterminism takes place relatively early in the process of deliberation, enabling the agent to perform genuine free actions. Zhu's paper faces the critical question that, even if determinism is false, the assumption that

a radical indeterminacy characterizes a decision-making process cannot secure a condition for rational, responsible free actions. After having introduced and replied to some major objections to libertarianism, Zhu provides an interesting account of how indeterminism can be considered a freedom-enhancing condition, arguing for what he calls a *deliberative libertarianism*. According to Zhu, indeterminacy, instead of being an obstacle to the libertarian's purposes, can be considered a crucial element of creativity that plays a critical role in practical deliberations and problem solving.

Three contributions from our call for papers conclude the Papers Section of the volume. They have been selected through a blind review process from among many other contributions we received. The first of them, by Liz Disley, emphasizes the role of social interactions in self-perception. The author focuses on the phenomenological experience of collective work as a paradigmatic example of intersubjectivity and human interaction. Following suggestions from Hegel, Husserl and Merleau-Ponty, Disley argues that the experience of physical work can improve one's own capacity for intersubjectivity, thus enhancing the role of the agent's embodied nature.

The second paper, by Susi Ferrarello, focuses on the notion of practical intentionality and investigates how it affects a decision-making process. Relying on a phenomenological approach, combining Husserl's theory of knowledge with Husserl's conception of will, the author defines a balance between logical and practical acts, showing how logical reason is necessary to give voice to our knowledge of reality, while practical reason is the starting point for every logical act.

Finally, David Vender's paper focuses on the role of acquired skills as emblematic aspects of action. According to the author, we do not have to be fully aware of our contribution to an action for it to count as a genuine act, nor do we necessitate a rational justification of it, but we must be able to adapt ourselves to the perceived situation. In view of that, Vender points out the critical role of balancing underlying perceptual and bodily orientation in executing complex actions.

As usual, we are also publishing a series of commentaries that provide new takes on well-established texts. They offer new, challenging arguments on the timeless questions concerning theory of agency and free will. Commentaries in this issue include the works of Roberta Lanfredini on Merleau-Ponty, Lorenzo Del Savio on Walter, Roberto Di Letizia on Wegner, Elisabetta Sirgiovanni on Libet, Freeman and Sutherland and, finally, Torrengo on Pereboom.

The volume also includes reviews of more recently published books that we are confident will provide arguments for discussion for many years to come. Among the many volumes published in the fields of theory of agency and free will, we selected the books by Laurence Shapiro, reviewed by Andrea Danielli, Sean Spence, reviewed by Roberto Di Letizia, Robert Rupert, reviewed by Mirko Farina, Alfred Mele, reviewed by Marco Fenici, Alva Noë, reviewed by Marco Spina, De Caro, Lavazza and Sartori, reviewed by Giuseppe Vicari, and Antony Chemero, reviewed by Silvano Zipoli Caiani.

Finally, the issue concludes with interviews of two prominent scholars: Sean Spence (interviewed by Duccio Manetti) and Daniel Dennett (interviewed by Marco Fenici and Stefano Di Piazza).

We would like to thank Livia Lentini and Alice Giuliani for their valuable assistance in editing this issue.

Dynamics, Agency and Intentional Action

Michael Silberstein[*]
silbermd@etown.edu

Anthony Chemero[**]
tony.chemero@fandm.edu

ABSTRACT

The complex systems approach to cognitive science invites a new understanding of extended cognitive systems. According to this understanding, extended cognitive systems are heterogenous, composed of brain, body, and niche, non-linearly coupled to one another. In our previous work, we have argued that this view of cognitive systems, as non-linearly coupled brain-body-niche systems, promises conceptual and methodological advances on a series of traditional philosophical problems concerning cognition, reductionism, and consciousness. In this paper, we discuss agency and intentional action in light of this view of cognition.

INTRODUCTION

Philosophical problems concerning intentional action, agency, volition and free will form a tangled knot. Just as with the hard problem of consciousness, most views on these problems tend to lead to dualism or eliminativism of one sort or another. For example, these views typically end with the idea that free will is either a force wielded by a homuncular agent or the idea that free will and agency are illusions. As many have noted, both sides tend to share Cartesian conception of self and action, more or less naturalized, and both sides tend to agree that reification of agency or its elimination are the only options. This conception includes the assumptions that action is caused by disembodied, internal representations (intentions, beliefs, desires, and reasons) wielded by

[*] Elizabethtown College
[**] Franklin & Marshall College

agents, all residing in the head. Intentions are understood as prior to actions and are detached from behavior. We reject all these assumptions in favor of a dynamical account of intentional action and agency; an account that allows us to avoid the extremes of dualism and eliminativism about intentional action and agency. However, unlike many other extended accounts of agency and action, we argue that extending agency and action makes them less susceptible to reification or elimination, not more. We are certainly not alone in trying to tell this story, see for example Juarrero 2009 and 2010, and a collection of articles devoted to a more embodied, embedded and extended account of intentional action and agency (Grammont *et al.* 2010).

We follow the strategy set out by Ryle in *Thinking and Saying* (1979). There, Ryle wants to describe thinking in a way that is not reductionist, but still avoids inflating thinking into something mysterious, because «Reductionist and Duplicationist theories are the heads and tails of one and the same mistake» (Ryle 1979, p. 80)

> The specific notion of Thinking, which is our long term concern, has been duly deflated by some philosophers into Nothing But such and such; and duly reinflated into Something Else as Well. (Ryle 1979, p. 80)

We do not endorse Ryle's story about thinking, but we do agree with his contention that the right story about it must be neither reductionist nor duplicationist. We think the same is true of agency and intentional action.

In previous work, we laid out a story about cognition and conscious experience that is neither reductionist nor duplicationist (Chemero 2009, Silberstein and Chemero, forthcoming). Consciousness and cognition are not Nothing But brain activity, but this does not mean they are to be reified as Something Else as Well. In this paper, we extend that approach to intentional action and agency. Our claims about action and agency are based on a particular conception of conscious cognitive agents that we call extended phenomenological-cognitive systems. The first part of the paper is devoted to characterizing that account and the second part will unpack the implications for intentional action and agency.

Extended Cognitive Systems

We have argued that, at least in some cases, cognitive systems are extended brain-body-environment systems (Chemero 2009; Silberstein and Chemero to

appear). We are not alone in defending what is now often called 'extended cognition'. But, as we will make clear below, our understanding of extended cognition is importantly different from most others. First, though, it is important to be clear on just what it is for cognition to be extended. To do so, consider a taxonomy offered by De Jaegher, Di Paolo and Gallagher (2010), concerning three ways in which features of the extra-bodily environment might be related to some cognitive phenomenon. First, the features might provide the *context* in which the cognitive phenomenon occurs, such that variations in the features produce variations in the cognitive phenomenon. Second, the features might *enable* the cognitive phenomenon, in the absence of the features, the cognitive phenomenon cannot occur. Third, the environmental features might be *constitutive parts* of the cognitive phenomenon. Only in this third case, when environmental features form constitutive parts of the cognitive phenomenon, is the cognitive system genuinely extended. (Note that De Jaegher *et al.* provide examples in which interpersonal social coordination plays each of these roles in social cognition, thus demonstrating that social cognition is at least sometimes extended.)

The empirical basis for our arguments that environmental features are sometimes constitutive parts of cognitive systems is research in dynamical modeling in cognitive science. Dynamical models have been used in psychology for at least 30 years (since Kugler *et al.* 1980), and have since then been employed with increasing frequency throughout neuroscience and the cognitive sciences. In dynamical systems explanation, one adopts the mathematical methods of non-linear dynamical systems theory, thus employing differential equations rather than computation as the primary explanatory tool. Dynamical systems theory is especially appropriate for explaining extended cognition because single dynamical systems can have parameters on each side of the skin. That is, we might explain the behavior of the agent in its environment over time as coupled dynamical systems, using something like the following coupled, non-linear toy equations, from Beer (1995, 1999):

$$\frac{dx_A}{dt} = A(x_A; S(x_E))$$

$$\frac{dx_E}{dt} = E(x_E; M(x_A))$$

where A and E are continuous-time dynamical systems, modeling the organism and its environment, respectively, and $S(x_E)$ and $M(x_A)$ are coupling functions from environmental variables to organismic parameters and from organismic variables to environmental parameters, respectively. Although in everyday conversation, we treat the organism and environment as separate, they are best thought of as comprising just one system, U. Rather than describing the way external (and internal) factors cause changes in the organism's behavior, such a model would explain the way U, the system as a whole, unfolds over time.

In those cases in which cognitive systems are best characterized as non-linearly coupled brain-body-environment systems that receive a dynamical explanation, the cognitive system is extended. When the constituents of a system are highly coherent, integrated, and correlated such that their behavior is a nonlinear function of one another, the system cannot be treated as truly a collection of uncoupled individual parts. Thus, if brain, body and environment are non-linearly coupled, their activity cannot be ultimately or best explained by decomposing them into sub-systems or system and background. Hence, they are one extended system, with brain, body and environmental features all serving as constitutive parts.

We can demonstrate this with an example. First, a little background: Work this decade has shown that 1/f noise (a.k.a., pink noise or fractal timing) is ubiquitous in smooth cognitive activity and indicates that the connections among the cognitive system's components are highly nonlinear (Ding *et al.* 2002; Riley and Turvey 2002; Van Orden *et al.* 2003, 2005; Holden *et al.* 2009). Research on the role of 1/f noise in cognition has allowed a new (and improved!) way to address some central issues in cognitive science, including allowing experimental approaches to questions that were thought to be "merely philosophical". [1] For example, Van Orden, Holden and Turvey (2003) use 1/f noise to gather direct evidence showing that, in certain cases, cognitive systems are not modular; rather these systems are fully embodied, and include aspects that extend to the periphery of the organism. Van Orden, Holden and Turvey (2003, 2005, 2009) argue that 1/f noise found in an inventory of cognitive tasks is a signature of a "softly assembled" system sustained by *interaction-dominant dynamics*, and not *component-dominant dynamics*. In component-dominant dynamics, behavior is the product of a

[1] See Stephen *et al.* 2009; Stephen and Dixon 2009; Dixon *et al.* to appear for some recent examples.

rigidly delineated architecture of components, each with pre-determined functions; in interaction-dominant dynamics, on the other hand, coordinated processes alter one another's dynamics, with complex interactions throughout the system. For example, when, as part of an experiment, a participant is repeating a word, a portion of her bodily and neural resources assemble themselves into a «word-naming device» (Van Orden *et al.* 2003, p. 346). Soft device assembly as the product of strongly nonlinear interactions within and across the temporal and spatial scales of elemental activity can account for the 1/f character of behavioral data, while assembly by virtue of components with predetermined roles and communication channels cannot. The key point for current purposes is that only when dynamics are component dominant is it possible to determine the contributions of the individual working parts to the overall operation of the system; in a system whose dynamics are interaction dominant, all of the system's parts are constitutive.

Finally, to the example: Dotov, Nie and Chemero (2010) describe experiments designed to induce and then temporarily disrupt an extended cognitive system, demonstrating that artifacts beyond the organism's periphery, can participate in the interaction-dominant dynamics of a human-tool system.

Participants in these experiments play a simple video game, controlling an object on a monitor using a mouse. At some point during the one-minute trial, the connection between the mouse and the object it controls is disrupted temporarily before returning to normal. Dotov *et al.* found 1/f noise at the hand-mouse interface while the mouse was operating normally, but not during the disruption. As discussed above, this indicates that, during normal operation, the computer mouse is part of the smoothly functioning interaction-dominant system engaged in the task; during the mouse perturbation, however, the 1/f noise at the hand-mouse interface disappears temporarily, indicating that the mouse is no longer part of the extended interaction dominant system. These experiments therefore were designed to detect, and did in fact detect, the presence of an extended cognitive system, one in which features of the environment are constitutive parts. The fact that such a mundane experimental setup (using a computer mouse to control an object on a monitor) generated an extended cognitive system suggests that extended cognitive systems are quite common. And note that because the system displayed interaction-dominant dynamics, it is not possible to separate any component of the system as playing essentially cognitive roles, while other

components are mere tools. We will return to this example repeatedly in this paper.

Extended Phenomenology-Cognition

In Chemero 2009 and, especially, Silberstein and Chemero to appear, we have argued that if features of the environment are sometimes constitutive parts of cognitive systems, it is attractive to view consciousness as being also partly constituted by features of the environment.[2] We claim that cognition and conscious experience are inseparable and therefore extended, and thus we often speak of 'extended phenomenological-cognitive systems'. In such systems, conscious experience is neither Nothing But brain activity, nor Something Else as Well (i.e., qualia). Because nothing in the claims we make about agency and action depends on the extension of conscious experience, we will not argue for extended consciousness in detail here. We will however use the phrases 'extended phenomenology-cognition' and 'extended phenomenological-cognitive systems'. We do so to differentiate our view from those of other proponents of extended cognition. One of the most important ways in which our view differs from others is that we embrace *antirepresentationalism*. In extended cognitive science, like the Dotov *et al.* experiments described above, non-linearly coupled animal-environment systems are shown to form just one unified, interaction-dominant system. The unity of such a system removes the pressure to treat one portion of the system as representing other portions of the system. Because the mouse and the object it controls on the monitor are constituent parts of the interaction-dominant cognitive system, there is no separation between the cognitive system and the environment that must be bridged by representations. So extended cognition invites antirepresentationalism. This antirepresentationalism is the key to the understanding extended cognitive systems as extended phenomenological-cognitive systems. As we will see below, it is also the key to the understanding of agency and action.

[2] See also Rockwell 2005.

Characterizing Extended Phenomenological-Cognitive Systems

We propose that extended phenomenology-cognition is to be understood as a variety of niche construction, one in which the constructed niche is an animal's cognitive and phenomenological niche. In biological niche construction, the activity of some organism alters, sometimes dramatically, its own ecological niche as well as those of other organisms (Olding-Smee *et al.* 2003). These animal-caused alterations to niches have profound and wide-reaching effects over evolutionary time. Phenomenological-cognitive niche construction has its effects over shorter time scales – an animal's activities alter the world as the animal experiences it, and these alterations to the phenomenological-cognitive niche, in turn, affect the animal's behavior and development of its abilities to perceive and act, which further alters the phenomenological-cognitive niche, and on and on.

Following enactive cognitive scientists (e.g., Maturana and Varela 1980; Thompson 2007; Di Paolo 2009) and ecological psychologists (e.g., Kelso *et al.* 1980; Swenson and Turvey 1991; Kelso 1995; Chemero 2008), we take animals and their nervous systems to be *self-organizing* systems. The animal's nervous system has an endogenous dynamics, which generates the neural assemblies that both compose the nervous system and constitute the animal's sensorimotor abilities. These sensorimotor abilities are the means by which the animal's niche couples with and modulates the dynamics of the animal's nervous system. These sensorimotor abilities are coupled with the niche, i.e., the network of affordances available to the animal (Gibson 1979). See *Figure 1*. This yields three (approximately) nested self-organizing systems, coupled to one another in different ways and at multiple time scales. Over behavioral time, the sensorimotor abilities cause the animal to act, and this action alters the layout of the affordances available, and the layout of affordances perturbs the sensorimotor coupling with the environment (causing, of course, transient changes to the dynamics of the nervous system, which changes the sensorimotor coupling, and so on). Over developmental time, the sensorimotor abilities, i.e., what the animal can do, determines what constitutes the animal's niche. That is, from all of the information available in the physical environment, the animal learns to attend to only that which specifies affordances complementing the animal's abilities. At the same time, the set of affordances available to the animal profoundly influence the development of the animal's sensorimotor abilities. So we have a three-part,

coupled, nonlinear dynamical system in which the nervous system partly determines and is partly determined by the sensorimotor abilities, which, in turn, partly determine and are partly determined by the affordances available to the animal. Also note that affordances and abilities are not just defined in terms of one another, but causally interact in real time and are causally dependent on one another in a nonlinear fashion.

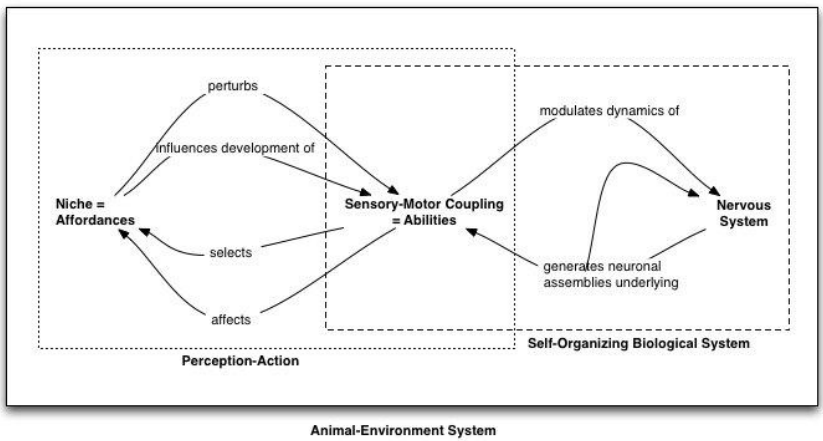


Figure 1

Understanding extended phenomenological-cognitive systems as genuinely phenomenological systems requires understanding affordances. Affordances are not independent properties of an animal's physical environment. They are irreducibly relational features of combined animal-environment systems, features that the animal perceives and uses to guide its action (Chemero 2003, Stoffregen 2003). The animal's behavioral niche, the set of affordances that it has learned to perceive and act upon, just is the environment as the animal experiences it. This underwrites a variety of *phenomenological realism*, or realism about the environment animals act in, think about, and consciously experience. Indeed, the entire system, including the environment as experienced, is required to account for and explain cognition. On this view, cognition and conscious experience are neither Nothing But brain activity, nor are they a dualistic Something Else as Well – they are the ongoing adaptive activity of the animal in its niche.

EXTENDED PHENOMENOLOGICAL-COGNITIVE SYSTEMS: PLASTICITY AND ROBUSTNESS

In order to more fully develop the idea that extended phenomenological-cognitive systems are multi-scale self-organizing systems, in this section we connect extended phenomenology-cognition to another recent topic in biology, the relationship between plasticity, robustness and autonomy in development.[3] Let us begin with phenotypic plasticity, wherein genetically identical individuals will frequently develop very different phenotypic traits when exposed to different environments or environmental conditions (Kaplan 2008). In general, a single genotype or genome can produce many different phenotypes depending on environmental and developmental contingencies. Phenotypic plasticity is just one example of the epigenomic processes in which various mechanisms create phenotypic variation without altering base-pair nucleotide gene sequences. These processes alter the expression of genes but not their sequence. In phenotypic plasticity, differential environmental conditions can lead to different phenotypic characteristics, but there are also cases where genetic or environmental changes have no phenotypic effect. *Robustness* is the persistence of a particular organism's traits across environmental or genetic changes. For example, in many knock-out experiments, a particular gene (or group of genes) known to be involved in the production of a protein or phenotypic trait is disabled, without disturbing the production of the protein or the development of the trait in question (Jablonka and Lamb 2005).

Together, plasticity and robustness imply that organismal processes have a fair measure of autonomy, in that organismal processes are maintained despite genetic and environmental disruptions. To account for the autonomy of the organism from both genetic and environmental changes, developmental biologists have called upon dynamical systems theory. The ongoing self-maintenance and development of an organism acts as a high-order constraint, which enslaves the components necessary to maintain its dynamics. Because of this, a developing system will have highly flexible boundaries, and will be composed of different enslaved components over time. This flexibility serves the autonomy of the developing organism, making it more likely to be viable. Autonomy is sometimes cashed out in terms of recursive self-maintenance.

[3] See also Thompson 2007.

That is, some systems are autonomous in that they can maintain stability not only within certain ranges of conditions, but also within certain ranges of changes of conditions: they can switch to deploying different processes depending on conditions in the environment.

The same is true, we believe, of extended phenomenological-cognitive systems. The coupled, dynamical phenomenological-cognitive system is highly opportunistic, encompassing different resources at different times. To use the language of dynamical systems theory once again, the extended phenomenological-cognitive system can be characterized as a set of order parameters that enslave components of brain, body and niche as needed in order to maintain itself. This means that the boundaries of the extended phenomenological-cognitive system will change (sometimes very rapidly) over time. And, as in the case of biological autonomy, the flexibility of the boundaries of extended phenomenological-cognitive systems is crucial to their self-maintenance. Autonomy as we are describing it here is the maintenance appropriate relations among the nervous system, the body and the environment, i.e., the maintenance of affordances and the cognitive-phenomenological niche. Thompson and Stapleton (2008) call this "sense-making".

> Organisms regulate their interactions with the world in such a way that they transform the world into a place of salience, meaning, and value – into an environment (Umwelt) in the proper biological sense of the term. This transformation of the world into an environment happens through the organism's sense-making activity. Sense-making is the interactional and relational side of autonomy. (Thompson and Stapleton 2008, p. 3)

This sense-making is the activity through which extended phenomenological-cognitive systems learn about, think about, and experience the world. Indeed, it is the activity through which they have a world.

Extended Phenomenological-Cognitive Systems: Action and Agency

Our view is that biological agents are best conceived as extended phenomenological-cognitive systems, and that extended phenomenological-cognitive systems engage in purposeful action. Indeed, it is better to say that the dynamical activity of extended phenomenological-cognitive systems *is* purposeful action. What are the consequences of this understanding of agency

and purposeful action? We begin by pointing out that there are significant areas of agreement between our position and that of others who advocate embodied, embedded and extended accounts of agency and action. We agree that agents are not just a sequence of decision making conscious states. We agree that one should endorse causal and explanatory pluralism (Chemero and Silberstein 2008) when it comes to explaining action. We agree that actions are processes extended in space and time, and that agents who engage in actions are extended in space and time and include aspects of the surrounding environment, social and physical, past and present, and perhaps even future (Clark 2007, p. 107). These are the points of agreement; where we differ from other proponents of extended agency is far more telling.

The first place we differ from Clark, and most other proponents of extended cognition, is over the role of computation in explaining cognition. Indeed, the debate about extended cognition is just an in house dispute over how wide computational processes are.[4] Extended phenomenological-cognitive systems do not function by representing the environment; the system and the environment are inseparable, so there is no need for intervening representation. On the conceptions of computation that have been used by cognitive scientists, computation requires representation (Fodor 1981). So extended phenomenological-cognitive systems are not computational systems; on our view, unlike many others who discuss extended cognition, cognition is not computation.

Moreover, the view of extended cognition as wide computationalism (Wilson 1995, 2004; Clark 1997, 2007) treats extended cognition as synonymous with *distributed* cognition. For example, in the ur-example of wide computation, the resources used to carry out long division are distributed among multiple separate components: a human brain, visual system, and motor system, along with the chalk and chalkboard on which the problem is written. The computational processing is distributed among these separate components, and the system like this would exhibit component-dominant dynamics as a whole. In contrast, extended phenomenological-cognitive systems are extended, but they are not distributed in the way Clark suggests. As we saw with the Dotov *et al.* study described above, the non-linear nature of extended phenomenological-cognitive systems, their robustness and their plasticity all imply that the systems are softly assembled, exhibiting and

[4] See the papers collected in Menary 2010.

sustained by interaction-dominant dynamics, and not component-dominant dynamics. The soft assembly is the product of strongly nonlinear interactions within and across the varying temporal and spatial scales of extended phenomenological-cognitive systems. It is driven by order-parameters in a higher-dimensional state space that both determine the expanding possibilities for the system as a whole and constrain the degrees of freedom of the more basic components in order to maintain the system as an autonomous, self-organizing unity. Because of (1) the time scale differences in the components' interactions and the dynamics of the whole system, and because (2) the same dynamics of the whole is often realized by multiple components (i.e., the system exhibits self-similarity at multiple spatial and temporal scales, which can be detected as 1/f noise), the system as a whole has a significant degree of autonomy from its components. The point is that extended phenomenological-cognitive systems are autonomous systems that are made up of components, but have dynamics that are not determined by the components (i.e., the dynamics are interaction dominant). This is in contrast with wide computational systems, which have component-dominant dynamics.

This difference between extended phenomenological-cognitive systems, which are extended but not distributed, and wide computational systems, which are distributed, is important to the discussions of agency and action. Taking cognition to be distributed, as it is in wide computational systems, makes agency ripe for elimination. Clark, for example, says

> what we really need to *reject,* I suggest, is the seductive idea that all these various neural and non-neural tools need a kind of stable, detached user. Instead, it is just *tools all the way* down. (Clark 2007, p. 111)

Clark also frames the debate in terms of the following dilemma: agency and action are just "tools all the way down" or they require a neural, functional center of consciousness, a central *self* relative to whom all neural, technological resources are mere tools (Clark 2007, p. 113). Clark is not alone in framing the state of play in this way. Ismael, for example, argues that we are forced between either a self-representation playing a causal role or mere input-driven self-organization; that is, real self-governance versus mere self-organization (Ismael 2010). The extended phenomenological-cognitive systems conception of agency and action shows that this is a false dilemma. The agency of extended phenomenological-cognitive systems is neither Nothing But tools nor Something Else as Well (a reified self-representation). Moreover,

because agency in extended phenomenological-cognitive systems inheres in a single (extended, but non-distributed) system with interaction-dominant dynamics, it is natural to claim that this system, as opposed its tools, is responsible for the action. The agency, like the system, might be extended, but it is not distributed.

An important question, though, is whether this sort of agency, which does without a Something Else as Well, is genuine agency. It is. Following Barandiaran, Di Paolo, and Rohde (2009), we take it that agency has three necessary components: the agent must be an identifiable individual; the agent must do something; and there must be norms governing what the agent does. We can see this by, once again, considering the Dotov *et al.* experiment. In the experiment, an extended phenomenological-cognitive system composed of (parts of) a person, a mouse, and computer display was brought into being and then temporarily disrupted. This system does compose an identifiable individual: the system as a whole behaved as an individual, as is indicated by its having measurable 1/f noise at the interface between the person and the mouse. This 1/f noise was a feature of the system as a whole, rather than a feature of any of its components. The system did something: the video game that was played had a goal state, and the extended phenomenological-cognitive system's activity was aimed at bringing that goal state into being. Finally, it was apparent whether the person-mouse-monitor system was successfully attaining the goal state, and when the mouse disruption made attaining that goal state difficult or impossible to achieve, the character of the system's activity changed such that the 1/f noise disappeared. That is, the system's activity was governed by norms, and the system's behavior changed when it was not achieving those norms. This extended phenomenological-cognitive system displays the necessary characteristics of genuine agency.

Extended Phenomenological-Cognitive Systems: Intentional Action

We have explained how extended phenomenological-cognitive systems can be agents, and can act purposefully. We have, so far, said nothing about how they might have *intentions* or act intentionally. In *intentional action*, an agent's intention is said to cause action. Given our goals, it is essential that intentional action be neither Nothing But behavior, nor Something Else as Well. So

intention must not be merely causally prior to the action but must somehow correspond to the intentional structuring of action, without being something over and above the action. The question that arises is how can physical processes instantiate intentional action of this sort? The outline of the correct answer to this question can be found in Juarrero's pioneering application of dynamical systems thinking to intentional action and agency (Juarrero 1999, 2009, 2010). Juarrero argues that beliefs, intentions, reasons, and the like are not the efficient causes of action. Instead, they act as context-sensitive constraints, and serve as final or formal causes of action. This is possible, she says, because «mental phenomena should be describable mathematically as neural attractors» (Juarrero 2010, p. 265). Intentions in particular are described as «higher-dimensional, neurologically embodied long-range attractors with emergent properties» (Juarrero 2010, p. 267). And more specifically, intentions are «soft-assembled context-sensitive constraints operating as control parameters» (Juarrero 2010, p. 268). These intentions constrain the activity of the system, so that the action comes about. Although Juarrero does not use the exact same language that we do, what she is describing is the activity of softly assembled, self-organizing systems that display interaction-dominant dynamics. Thus we agree with Juarrero that intentions are best understood as control parameters, which are both composed of the system's components and also act as constraints on the activity of those components. This allows intentions to play a role in the generation of action without being identical to the system components and without being anything over and above the system.

There are, however, important differences between our view and Juarrero's. First, while she does stress that the environment gets folded into cognitive processes that are not just in the brain (Juarrero 2010, p. 265), we think that she is too closely focused on the brain and "self-organized neural states". Second, Juarrero's view is representationalist:

> A self-organized neural state is representational and symbolic if its central features are given not by the configuration's intrinsic physical properties but by the *information* it carries. (Juarrero 2010, p. 264)

Finally, and most importantly, we worry that Juarrero's view leans too much toward the elimination of intentions. The second and third of these differences in approach stem from the first. Because Juarrero takes intentions to be self-organizing neural processes, the parameters that govern their organization are

independent of the environment and the rest of the body. (That, after all, is what it is to be *self*-organizing.) Because of their independence from the environment and the rest of the body, there is pressure to treat them as representing the body and environment. And, given Juarrero's laudable wish to avoid reifying intentions and other mental entities, she ends up explaining the connection between these context-sensitive neural constraints and action by a body in an environment in a highly deflationary fashion, suggesting classical probability theory as a good analogy for such constraints (Juarrero 2010, p. 260). On this analogy, your intention to get a cup of coffee right now impacts your action the same way that the laws of probability influence the outcome of a coin toss. This strikes us as going beyond Nothing But, all the way to Nothing.

The solution here, of course, is to reject Juarrero's neural focus by taking intentions not to be self-organizing neural attractors that constrain the activity of the body, but rather to be order parameters of self-organizing extended phenomenological-cognitive systems that act as constraints on components of extended phenomenological-cognitive systems. This is perfectly in line with the way that constraints are discussed in physics, allowing them to be kinematic, geometric, or topological constraints, including various kinds of symmetries, or even boundary conditions. These constraints are features of a system that can impact the behavior of the system, and whether one wishes to call this impact formal cause or final cause, it is above all lawful and dynamical. These non-reified intentions genuinely constrain the activity of system without being something outside it.

Conclusion

We have been making the case that agents are extended phenomenological-cognitive systems, composed of a changing collection of components of the brain, body and niche. These systems exhibit interaction-dominant dynamics, so it is impossible to separate out the contributions from individual system components; this means they are extended, but not distributed. These systems are genuinely agents and engage in intentional action. Their intentions are order parameters that constrain the activity of system components, but do not act as efficient causes. These agents do not pop into existence (emerge) from complex brain dynamics, already armed with powers of intentionality and will. Rather, agent and environment are co-dependent sides of the same coin. In

other words, sense-making and agency go hand in hand. It is built into this conception of things that cognitive agents consciously experience the world in terms of their abilities and goals. Given this, there is no special mystery of how meaningful behavior could be possible. We are extended phenomenological-cognitive systems, which is to say that we are not brains in vats in representation-mediated contact with the environment we want to act in, somehow; instead, we are meaningful action.

REFERENCES

Barandiaran, X., Di Paolo, E., & Rohde, M. (2009). Defining agency. *Adaptive Behavior, 17*(5), 367-386.

Beer, R. (1999). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences, 4*(3), 91-99.

Beer, R. (1995). Computational and dynamical languages for autonomous agents. In R. F. Port & T. Van Gelder (Eds.), *Mind as Motion*, (pp. 121-147). Cambridge, MA: MIT Press.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology, 15*(2), 181-195.

Chemero, A. (2008). Self-organization, writ large. *Ecological Psychology, 20*(3), 257-269.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Chemero, A., &, Silberstein, M. (2008a). After the philosophy of mind. *Philosophy of Science, 75*(1), 1-27.

Clark, A. (1997). *Being There*. Cambridge, MA: MIT Press.

Clark, A. (2007). Soft selves and ecological control. In D. Ross, D. Spurrett, H. Kincaid & G. Lynn Stephens (Eds.), *Distributed Cognition and the Will: Individual Volition and Social Context*, (pp. 101-122). Cambridge, MA: MIT Press.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*(10), 441-447.

Ding, M., Chen, Y., & Kelso, J. A. S. (2002). Statistical analysis of timing errors. *Brain and Cognition, 48*(1), 98-106.

Di Paolo, E. (2009). Extended life. *Topoi*, *28*(1), 9-21.

Dixon, J., Holden, J., Mirman, D., & Stephen, D. (to appear). Multi-fractal development of cognitive structure. *Topics in Cognitive Science*.

Dotov, D., Nie, L., & Chemero, A. (2010). A demostration of the transition from readiness-to-hand to unreadiness-to-hand. *PLoS ONE, 5*(3), e9433.

Fodor, J. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Grammont, F., Legrand, D., & Livet, P. (Eds.) (2010). *Naturalizing Intention in Action*. Cambridge, MA: MIT Press.

Holden, J., Van Orden, G., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological Review, 116*(2), 318-342.

Ismael, J. (2010). Self-organization and self-governance. *Philosophy of the Social Sciences*. Published online, doi:10.1177/0048393110363435

Jablonka, E., & Lamb, M. (2005). *Evolution in Four Dimensions Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT Press.

Juarrero, A. (2009). Top-Down causation and autonomy in complex systems. In N. Murphy, G. F. R. Ellis & T. O'Connor (Eds.), *Downward Causation and the Neurobiology of Free Will*, (pp. 83-102). New York: Springer-Verlag.

Juarrero, A. (2010). Intentions as complex dynamical attractors. In J. H. Aguilar & A. A. Buckareff (Eds.), *Causing Human Actions: New*

Perspectives on the Causal Theory of Action. Aguilar. Cambridge, MA: MIT Press.

Kaplan, J. (2008). Review of *Genes in Development: Rereading the Molecular Paradigm* edited by Eva M. Neumann-Held and Christoph Rehmann-Sutter. *Biological Theory, 2*, 427-429.

Kelso, J. A. S. (1995). *Dynamic Patterns*. Cambridge, MA: MIT Press.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1980). Coordinative structures as dissipative structures I. Theoretical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior*. Amsterdam: North Holland.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. (Boston Studies in the Philosophy of Science, vol. 42). Dordrecht: D. Reidel Publishing Company.

Menrary, R. (Ed.) (2010). *The Extended Mind*. Cambridge, MA: MIT Press.

Riley, M., & Turvey, M. (2002). Variability of determinism in motor behavior. *Journal of Motor Behavior, 34*(2), 99-125.

Rockwell, T. (2005). *Neither Brain nor Ghost*. Cambridge, MA: MIT Press.

Ryle, G., & Kolenda, K. (1979). *On Thinking*. Oxford: Blackwell.

Silberstein, M, & Chemero, A. (to appear). Complexity and Extended Phenomenological-Cognitive Systems. *Topics in Cognitive Science*.

Stephen, D. G., & Dixon, J. A. (2009). The self-organization of insight: Entropy and power laws in problem solving. *Journal of Problem Solving, 2*(1), 72-102.

Stephen, D., Dixon, J, & Isenhower, R. (2009). Dynamics of representational change: Entropy, action, and cognition. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1811-1822.

Stoffregen, T. A. (2003). Affordances as properties of the animal-environment system. *Ecological Psychology, 15*(2), 115-134.

Swenson, R., & Turvey, M. (1991). Thermodynamic reasons for perception-action cycles. *Ecological Psychology*, *3*(4), 317-348.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Harvard University Press.

Thompson, E., & Stapleton, M. (2008). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi, 28*(1), 23-30. Published online, doi: 10.1007/s11245-008-9043-2.

Van Orden, G., Holden, J., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*(3), 331-351.

Van Orden, G., Holden, J., & Turvey, M. (2005). Human cognition and 1/f scaling. *Journal of Experimental Psychology: General, 134*(1), 117-123.

Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind.* Cambridge, MA: MIT Press.

Wilson, R. (2004). *Boundaries of the Mind.* Cambridge, MA: Cambridge University Press.

An Inter-Enactive Approach to Agency: Participatory Sense-Making, Dynamics, and Sociality[*]

Steve Torrance * *
stevet@sussex.ac.uk

Tom Froese * * *
t.froese@gmail.com

ABSTRACT

An inter-enactive approach to agency holds that the behaviour of agents in a social situation unfolds not only according to their individual abilities and goals, but also according to the conditions and constraints imposed by the autonomous dynamics of the interaction process itself. We illustrate this position with examples drawn from phenomenological observations and dynamical systems models. On the basis of these examples we discuss some of the implications of this inter-enactive approach to agency for our understanding of social phenomena in a broader sense, and how the inter-enactive account provided here has to be taken alongside a theory of larger-scale social processes.

1. INTRODUCTION

It is now two decades since the emergence of Enactivism as a distinctive approach within Cognitive Science, with the publication of *The Embodied*

* The authors wish to express their gratitude to a number of people for conversations from which they derived great benefit in the writing of this paper. These include Giovanna Colombetti, Stephen Cowley, Hanne De Jaegher, Ezequiel Di Paolo, Madeline Drake, Shaun Gallagher, John Stewart; colleagues at the Life and Mind seminars at the University of Sussex; and participants at a workshop on Enaction and Social Cognition, Battle, Sussex, UK, in September 2008. They would also like to acknowledge the support of the EUCognition network, which funded their participation at the above workshop. Froese's research is funded by a postdoctoral fellowship of the Japanese Society for the Promotion of Science (JSPS).

* * Centre for Research in Cognitive Science, Department of Informatics – University of Sussex

* * * Ikegami Laboratory, Department of General Systems Studies – University of Tokyo

Mind (Varela *et al.* 1991). This volume was a broadside against traditional conceptions of mind and agency – in particular the dominating notion that a cognitive agent's interactions with the world are essentially mediated by an internal information-processing device, epitomized by the digital computer, linked to sensors and effectors. On this picture the agent's brain receives sensory inputs which enables the brain's information-processing routines to update its internal model of the world, modify its action-plans and generate executive commands to effect physical changes in the world – what Rodney Brooks (1991) described as the 'sense-model-plan-act' view of cognition and agency. At the time of The Embodied Mind the dominant question being asked was: is the information processing device to be thought of primarily in terms of symbolic AI models or in terms of some form of connectionist architecture. Varela and colleagues were inspirational in offering 'enactivism' as a new departure from these 'internalist' models, and many other novel approaches appeared, calling themselves, variously, 'embodied', 'embedded', 'dynamic', and so on.

In fact a bewildering number of different proposals were made under the 'enactivist' banner[1] – so that it is rather difficult to find a concise summary of what enactivism, in essence, was proposing. However the enactive approach to cognition and agency can be broadly summarized in terms of five interlocking themes. See Thompson 2005, 2007; Torrance 2005, which address the foundational question: What is it to be a (cognizing, conscious) agent? The five-fold response is as follows: it is (a) to be a biologically autonomous (autopoietic) organism – a precarious, far-from-equilibrium, self-maintaining dynamic system; (b) with a nervous system that works as an organizationally closed network, whose function is to generate *significance* or *meaning*, rather than (as in the 'sense-model-plan-act' model) to act via a set of continually updated internal representations of the external world; (c) the agent's sense-making arises in virtue of the its dynamic sensorimotor coupling with its environment, such that (d) a world of significances is 'enacted' or 'brought forth' by a process whereby the enacted world and the organism mutually co-determine each other; and (e) the experiential awareness of that organism arises from its lived embodiment in the world.

These five themes draw upon a number of theoretical traditions, for example, the autopoietic theory of Maturana and Varela (1987), the

[1] See Torrance 2005 for a catalogue of some of these.

phenomenology of Merleau-Ponty (1945) and recent work on dynamical systems (e.g., Port and Van Gelder 1995), as well as (in some interpretations of enactivism) also leaning heavily on themes from Eastern mindfulness meditation traditions (stressed in particular in Varela *et al.* 1991). Putting all these various strands together we have a view of agency which stresses how an agent and the world in which that agent acts can, in an important sense, be seen as 'co-constituting' or 'co-enabling' one another. The enactive approach to cognitive science has come a long way since it was first initiated by Varela and colleagues, as demonstrated, for instance, by the subtle and extended treatment in Thompson (2007). An initial focus on the embodied phenomenology and sensorimotor dynamics of perception (O'Regan and Noë 2001, Noë 2004)[2] has come to be complemented by a renewed interest in biological autonomy (putting more emphasis, for example, on autopoiesis)[3], and this resulted in a sharpened conception of sense-making in relation to autopoiesis, a stress on the importance of adaptivity (Di Paolo 2005), and more recently, an enactively-focused characterization of agency (Thompson 2005, Barandiaran *et al.* 2009). Other recent treatments of enactivist themes include collected papers on intersubjectivity, empathy and sociality (Thompson 2001, Di Paolo 2009); on enactive experience (Torrance 2005, 2007); on autonomy (Barandarian and Ruiz-Mirazo 2008); on enactivism in relation to other post-cognitivist views of mind (Kiverstein and Clark 2009, Menary 2010); and on enactivism as a new paradigm for the cognitive sciences (Di Paolo *et al.* 2010).

During the progression of this extended discussion it has become evident that it is not just the internalism and representationalist nature of classical cognitivism that has to be challenged, but also a 'methodological individualist' or 'methodological solipsist' approach to cognition and agency.[4] This individualistic picture has been challenged in many ways by enactivist and other perspectives which stress embodied and embedded, features of agency and cognition. Thus there has been a growing focus on the intersubjective, or interactive, nature aspects of experience, knowledge and agency.[5] De Jaegher

[2] See also Torrance 2002.

[3] See Varela 1997, Weber and Varela 2002.

[4] See Fodor 1980 for a classical defence of an account of cognition which explicitly takes this character, under the label 'methodological solipsism'.

[5] This emphasis on interaction and intersubjectivity was given an important impetus by an emphasis on second-person methods for investigating consciousness, and empathy as a central feature

and Di Paolo's (2007) enactive account of social interaction provided a new departure by introducing the concept of *participatory sense-making*. Their account, which drew inspiration from autopoietic accounts of biological autonomy (e.g., Maturana and Varela 1987), and from research in artificial life and evolutionary robotics, proposed that inter-individual interaction processes can take on an autonomous organization of their own.

De Jaegher and Di Paolo's paper is a key study in a number of recent works which mark a growing interest in the role of the interaction between agents for understanding the nature of agency.[6] For instance, it has been argued that the inter-individual interaction process can constitutively shape forms of individual agency (De Jaegher and Froese 2009), that inter-agent interaction is a necessary condition for the shift from minimal to 'higher-level' cognition (Froese and Di Paolo 2009), and that historically based impersonal norms an essential background in human social agency (Steiner and Stewart 2009; see also below section 4). In addition, Di Paolo, Rohde and De Jaegher (2010) have investigated the importance of an enactive account of social cognition to understanding the nature of play. Moreover, the idea that interactive processes are defined by a certain autonomy which both conditions and is conditioned by the autonomy of the interacting individuals has profound repercussions for our understanding of emotion, values and ethics (Colombetti and Torrance 2009).

In what follows we will further substantiate the idea that many kinds of agency, in particular the agency of human beings, cannot be understood separately from understanding the nature of the interaction that occurs between agents. We begin with a discussion of some illustrative examples, drawn from common experience, that show how the relative autonomy of the interactive process itself can sometimes facilitate and sometimes hinder our individual goals. However, the majority of cognitive scientists working on interpersonal interaction and social cognition are likely to remain unconvinced that these examples show that inter-individual interaction processes can indeed play a constitutive role in determining the character of individual agency. Accordingly, we discuss a series of simulation models which serve as concrete proof of concepts, and which enable us to analyze the dynamics of the interaction process in a precise mathematical manner. Effectively, the models help us to demonstrate that it is possible to treat an inter-individual interaction

of this emphasis, as seen in the collection of papers which was published shortly after Varela's death (Thompson 2001). See also, for example, Gallagher 2005.

[6] See also Di Paolo 2009.

process as one whole dynamical system, and that this global system has properties that modulate the flow of component activity and yet cannot be reduced to the activity of any of the individual components. Having put the enactive approach to social interaction on a more solid footing, we proceed to discuss some of the implications this change in perspective has for our understanding of social processes in a sense broader than that which is limited to the inter-individual real-time interactions which are the major focus in the earlier part of our discussions.

2. When Interacting with Others Takes on a Life of Its Own

One of the key ideas to have emerged from the 'participatory sense-making' literature is that the unfolding of an interaction between two or more people has an autonomy of its own which is separate from the autonomy of the individual participants. This idea will be given more theoretical weight later on, but here it is illustrated intuitively. The relative autonomy of an inter-individual interaction process may be encountered as an organizing (enabling and constraining) influence on the unfolding events of the interaction from the perspective of the interacting individual agents. Depending on the circumstances, the autonomous organization of the interaction process itself can facilitate or hinder the realization of the autonomous goals of the agents. Here we illustrate these two types of situation by drawing on some concrete examples from common experience, at first intuitively or pre-theoretically described.

2.1 How the interaction process facilitates individual actions

To begin with, it may be easiest to illustrate the constitutive role of the interaction process for individual actions by recalling a social situation in which we were engrossed in a conversation. It can happen that the flow of the interaction carries us along quite effortlessly, with every one of our actions prompting our interlocutor to respond with a complementary reaction, which in turn evokes another response from us, and so forth, back and forth. In this way we can describe the conversation as a stable social situation because of the mutually reinforcing actions of the interlocutors.

At the same time we can also look at the role of the conversation itself. In other words, the fact that we are situated in an engaging conversation means

that in response to what the other has said we are more likely to say or do something appropriately engaging in turn, and the fact that we are more likely to respond in this way also means that our actions are ensuring that we continue to be situated in an engaging conversation. We are thus faced with a self-perpetuating social interaction process, whereby the conversational nature of the situation co-constitutes the individual's gestures, and the individual's gestures co-constitute the conversational nature of the situation.

But does it really make sense to give a co-constitutive role to the structure of the inter-individual interaction process itself? How do we know that we are not simply dealing with the linear sum of the individual gestures? In the case of a social situation in which the individual goals of the partners are mutually reinforcing, it is indeed difficult to assess whether we need to appeal to any additional interactional process at the inter-individual level in order to explain what is going on. However, what about social situations in which the goals of the individual interactors are not aligned with the self-perpetuating structure of the interaction process?

2.2 How the interaction process hinders individual actions

De Jaegher and Di Paolo (2007, 2008) nicely illustrate this possibility by pointing out that verbal arguments are often self-perpetuating even despite the best intentions of those involved. In such cases every attempt to end the conflict by one or the other of the individuals may, in virtue of the inter-individual situation, provoke a response from the partner and will therefore, in spite of the individual's original goal, inadvertently give support to the continuation of the overall argumentative situation. Anyone who has experienced being entangled in such a self-perpetuating social conflict knows the feeling of being helpless to stop what turns out to be an inevitable continuation of the argument. On the personal level it can feel like what one is saying is somehow twisted in the interaction so that it comes out wrong, is misinterpreted, or simply remains ineffective.

The self-perpetuating verbal argument that no one wants to continue having is a rather extreme example, but self-sustaining interactions in which the structure of the social situation cannot be reduced to the sum of the individual actions are actually quite common in our daily lives. As a paradigmatic non-verbal example, De Jaegher and Di Paolo (2007) refer to the situation in which we encounter someone while walking along a corridor and we step aside to make way. It sometimes happens that we both step aside in the

same direction and are thus faced by the same impasse once again, which leads us to make another synchronous sideways step together, and so forth, until one of us finally makes a concerted effort to break the undesired interaction process and lets the other pass.

A common verbal example that we can all relate to is trailing conversations that we have difficulty in terminating. This can happen for instance when trying to end a phone call in a polite manner, such that every 'bye' and 'thanks' and 'see you soon' uttered by one of the speakers is followed by a complementary response by the other speaker, which then calls forth another response from the first speaker, and so forth. In this way the 'end' of the conversation continues because the social situation as such facilitates the exchange of mutually contingent responses, as well as because the cultural norms of our society make it difficult to simply hang up on someone who is still speaking. Accordingly, even though both callers may have the personal goal of terminating the call, they can find themselves unable to easily do so because additional responses are facilitated by the interactional nature of the social situation and the cultural constraint of not hanging up prematurely.

We will return to a fuller discussion of the role of cultural norms in shaping individual actions in a later section of this paper, but for now we simply want to highlight the fact that even the most basic inter-individual interaction processes can become self-perpetuating, autonomous structures in their own right, and that these relational structures can play a constitutive role for the enaction of individual actions (De Jaegher and Froese 2009).

2.3 How to avoid falling into social mysterianism

However, as Boden (2006) has correctly pointed out, this kind of approach to social interaction confronts us with a fundamental problem: how can we leave methodological individualism, which is still prevalent in the cognitive sciences, behind us without at the same time descending into some kind of social mysterianism? How can we scientifically grasp the notion that an inter-individual interaction process is not just *constituted by* the actions of several interacting individuals, but that this whole interaction process itself is also *constitutive of* the actions of those individuals as well?

Mainstream approaches to social cognition are ill equipped to address this important challenge because they remain narrowly focused on the cognitive

abilities of the brains of isolated individuals (usually characterized in terms of either theory theory or simulation theory).[7] Fortunately, however, these cognitivist and individualist approaches to social interactions are no longer the only game in town. In what follows we will show that it is possible to systematically study the nature of social situations, including their co-constitutive impact, by making use of some minimalist technological tools. Not only does this address the worry that an acceptance of the co-constitutive role of interaction processes and a rejection of theory of mind approaches leads to a non-scientific mysterianism about sociality; on the contrary, because our framework is based on the mathematics of dynamical systems theory, we are grounding the discussion in concrete models that are open to precise analysis.[8]

3. Dynamical Models of Inter-Enaction

In the previous section we have described two distinct types of social situation in which interactions between two or more individual agents can form a self-perpetuating dynamic structure at the level of the interactions themselves. In these types of social situation the interaction process entrains the actions of the individual interactors in such a way that they support the continuation of the interaction process itself. And, depending on the organization of the interaction process, it can either facilitate or hinder the realization of the individual interactors' goals accordingly.

3.1 Methodological considerations

However, it is not enough to simply describe these social situations in a

[7] For some recent critical accounts of conventional accounts of social cognition and 'mind-reading', see Gallagher 2001, 2005, 2008; Gallagher and Zahavi 2008, pp. 171-197. Dan Hutto's Narrative Practice Hypothesis provides a particularly fertile source for criticisms of orthodox approaches to social cognition (Hutto 2004, 2007; see also Gallagher and Hutto 2008).

[8] Note that we are also avoiding the category mistake committed by Theory of Mind approaches, which attempt to devise scientific explanations of social cognition by re-describing our personal-level abilities as hypothetical brain-based mechanisms (e.g., our personal-level ability to imagine ourselves in someone else's place makes a reappearance in the supposedly sub-personal simulation capacity of so-called mirror neurons). The enactive approach, on the other hand, does not have to make use of homuncular discourse when explaining sub-personal mechanisms underlying social interactions in terms of dynamical systems.

narrative manner and to affirm in an intuitive way the personal sense of being enabled or constrained in order to establish a science of social or interactive situations. What is additionally required is a basic proof of concept, which demonstrates that these narrative and phenomenological descriptions of the efficacy of interaction processes in certain social situations are not merely metaphorical embellishments of what is essentially a sum of individual actions. To show that the social interaction process itself can play a constitutive role for the actions of individual agents, we need to be able to show this process at work in a concrete model that allows for systematic exploration of the essential parameters.

One suitable way of satisfying this additional requirement is to take advantage of recent work in Evolutionary Robotics modeling. Since its beginnings in the early 1990s, the Evolutionary Robotics approach has established itself as a viable methodology for optimizing dynamical controllers for physical robots (Nolfi and Floreano 2000), as well as for synthesizing simulation models of what has become known as 'minimally cognitive behavior' (Beer 1996). The idea here is to set up an evolutionary algorithm that can automatically shape the dynamical system of a model agent so that it performs a given task in the simplest possible way, while still raising issues that are of genuine interest to cognitive scientists (Beer 2003, Harvey *et al.* 2005, Froese and Ziemke 2009). In the last decade there has been a growing interest in using this kind of Evolutionary Robotics approach to investigate the dynamics of social interactions (e.g., Iizuka and Di Paolo 2007a, Froese and Di Paolo 2008, Di Paolo *et al.* 2008), and the methodology has accordingly been extended to include 'minimally social behavior' as well (Froese and Di Paolo in press).

We will draw on some of our own modeling work in this area (Froese and Di Paolo 2010, in press), because the specific aim of these models is to serve as proof of concepts which demonstrate that interaction processes themselves can play a constitutive role in shaping individual actions over and above the sum power of the individual agents. In fact, as we have already argued extensively elsewhere (Froese and Gallagher 2010), the methodology of Evolutionary Robotics is well suited to complement phenomenological investigations in the cognitive sciences.

3.2 The dynamics of perceptual crossing

Froese and Di Paolo (2010) used an Evolutionary Robotics approach to

generate a series of agent-based simulation models whose minimalist task-design is directly based on a psychological experiment on perceptual crossing by Auvray, Lenay and Stewart (2009). The term 'perceptual crossing' denotes social situations in which the perceptual activities of two agents interact with each other (e.g., mutual touch or catching another's eye). Essentially, the study by Auvray and colleagues is an exploration of the most basic conditions that are necessary for participants to recognize each other by means of minimal technologically mediated interaction in a shared virtual space. Since this study is the original inspiration for the simulation models, we will describe it in a bit more detail first.

A schematic illustration of the overall experimental setup is shown in *Figure 1*. Two adult participants, acting under the same conditions, can move a cursor left and right along a shared one-dimensional virtual 'tape' that wraps around itself. They are asked to indicate the presence of the other's cursor-driven virtual 'body' by clicking a mouse button. The participants are in separate rooms and can only sense a tactile stimulation (on/off) on their finger, depending on whether the location of their cursor coincides with another object in the virtual space. Apart from each other's cursor object, participants can encounter a static object on the tape, or a mobile 'shadow' object that is fixed at a distance to the partner's cursor. All objects are strictly identical in size, and the two mobile objects (the other's cursor-driven 'body' and its attached 'shadow') perform identical movements. Importantly, only the other's cursor can be responsive to one's own movements since it provides tactile feedback to the other participant.

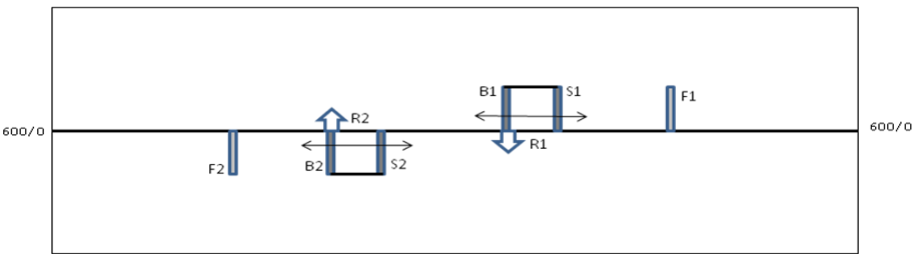


Figure 1: Visual schematic of the experimental setup of Auvray, Lenay and Stewart's (2009) study of perceptual crossing (adapted from Froese and Di Paolo 2010). Two participants inhabit a virtual space consisting of a 600 unit long 1-D toroidal (wrap-around) environment. The space is divided into two regions, 'Up' and 'Down'. Each region contains three objects,

shown as grey oblongs. These are (a) the participant's mouse-driven 'body' or 'avatar' (labeled B1, B2) which the participant can move left or right at will; (b) the body's 'shadow' (S1, S2) which moves in lockstep with the avatar; (c) a fixed object (F1, F2). In addition, each participant's has a receptor field (R1, R2 – shown as white arrows pointing into the other region), which move with the avatar's body, and which can overlap with each of the three objects in the other region as it moves left and right. In the actual experiment participants are blindfolded, and use a mouse to control the movement of their avatar; the other hand is placed on a custom-built tactile feedback device which issues an identical short vibration when the receptor field encounters an object in the opposing space.

Since all virtual objects are of the same size and only generate an all-or-nothing tactile response, the only way to differentiate between them is through the interaction dynamics that they afford. And, indeed, an analysis of the results revealed that, although they were often not aware of this fact, the participants did manage to locate each other successfully. Essentially, the reason for this success is that the ongoing mutual interaction afforded the most stable situation under these circumstances. If one participant's receptor field coincided with the other's body, thus activating the tactile feedback, the other participant's receptor field would simultaneously also coincide with the first participant's body, thus activating their tactile feedback, too. Accordingly, both participants were mutually engaged in the same interaction and neither of them had reason to disengage and to continue searching elsewhere. But if a participant happened to interact with the other's mobile shadow object (whose movements are an exact copy of the other's movements), the other would not receive any tactile feedback from their engagement and would continue searching, thus dragging their shadow object with them and terminating the other's attempt at interaction. Interaction with the shadow object is therefore inherently unstable, while mutual perceptual crossing is relatively stable. To be sure, interacting with a static object is stable too, but the lack of social contingency is given away when the interaction becomes too predictable, after a few iterations at least.

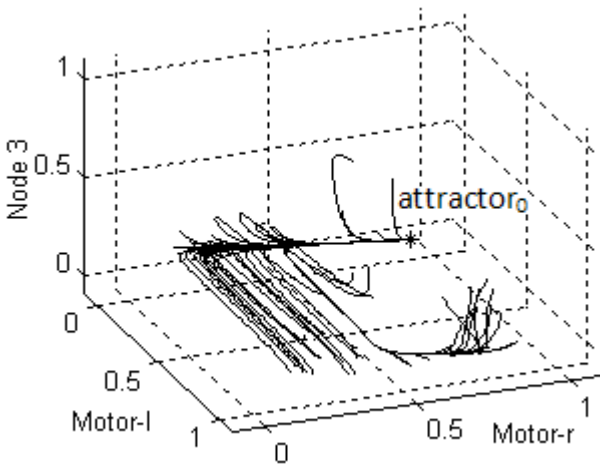
A closer look at the results reveals a special role of the interaction process in the overall outcome of the experiments. Interestingly, the participants 'failed' to achieve the task individually, because there was no significant difference between the probability of a clicking response to the other's body and the other's shadow object (Auvray *et al.* 2009, p. 39). In other words, on an interaction to interaction basis, the participants were unable to distinguish

between those situations that were characterized by social contingency and those that weren't. However, they still managed to solve the task collectively because of the self-sustaining dynamics of the interaction process. That is, *at the end of a whole trial the most clicks in total occurred during situations of actual perceptual crossing*. The upshot of these experiments therefore is that, even though it is impossible to distinguish the active partner from her irresponsive copy on an individual basis, it turns out that most clicks are made correctly because a mutual interaction is more likely to persist and participants are therefore more prone to face each other once again.

The value of modeling this psychological experiment has already been shown by Di Paolo, Rohde and Iizuka (2008), who used an Evolutionary Robotics approach to generate an agent-based simulation model which successfully replicated the main results of the study. At the same time it helped them in gaining some additional insights into the dynamics of the interaction process. For example, the problems that their model agents had with avoiding interactions with their respective static objects led them to predict similar difficulties for human participants. This prediction was already supported by the empirical data presented by Auvray and colleagues, but it had previously gone unnoticed.

Froese and Di Paolo (2010) continued this modeling research with the aim of gaining a better appreciation of the further potential of this general experimental setup and, at the same time, of improving our understanding of the constitutive role of the interaction process for individual behavior and agency. They began by using a similar modeling setup as that used by Di Paolo and colleagues (2008), and provided a comprehensive analysis of the evolved behavioral strategy by means of a set of simulated psycho-physical tests. The results of the original study and its first model were successfully replicated. The novel aspect of Froese and Di Paolo's re-implementation is the great simplicity of the 'neural' system of the evolved agents, which enables a detailed dynamical understanding of their behavior. An example of the kind of analysis that is made possible by this kind of model is shown in *Figure 2*.

(a)


Node 3
1
0.5
0
attractor0
0
0.5
Motor-l
1
0
0.5
Motor-r
1

(b)

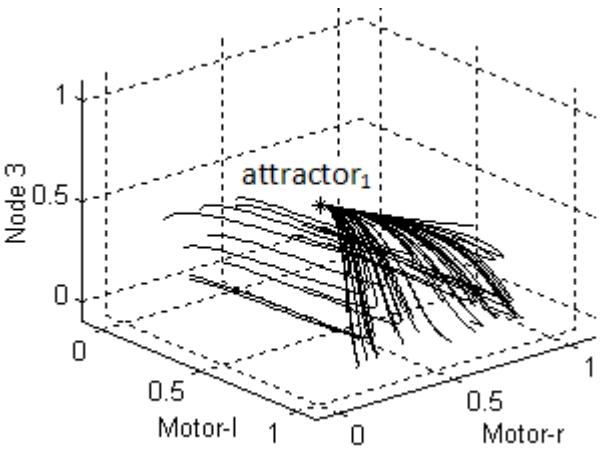

Node 3
1
0.5
0
attractor1
0
0.5
Motor-l
1
0
0.5
Motor-r
1

Figure 2: Illustration of the 3D state-space attractor landscape for the three-node continuous-time recurrent neural network (CTRNN) controlling the movements of one model agent (figure adapted from Froese 2009, p. 169). All nodes receive input from the agent′s receptor field and the output of the nodes labeled 'motor-r' and 'motor-l' determine the agent′s rightward and leftward velocity, respectively. Note that depending on whether the agent's receptor field is turned off (a) or on (b), the position of the point attractor, represented by a *, changes to a different region of state-space. The lines converging on the attractors represent a sample of possible state trajectories of the neural network (for 50 times the states of the network's nodes were initialized to random values drawn from a representative trial run and the network was allowed to settle for 8000 time steps). See text and Froese and Di Paolo 2010 for details.

When the continuous-time recurrent neural network (CTRNN) that controls movement of a model agent is decoupled from the 1-D environment, it is characterized by two fixed point attractors, (labeled attractor_0 and attractor_1) depending on whether the receptor field input is off or on. It turns out that the velocity of the agents is strongly coupled to the value of this parameter. This is indeed the basis for a tight sensorimotor coupling: the state of the receptor field input parameter is largely determined by the current movement of the agent in relation to its current environment (including potentially its relation to the other agent), and at the same time its current movement is largely determined by the state of the input parameter.

But this tight coupling should not be misunderstood as the mark of a purely reactive system, since the sensorimotor loop is mediated by a dynamical system with feedback connections. Moreover, because the contact sensor switches an agent's neural system between the two different attractor landscapes (with attractor0 and attractor1), the inter-individual interaction process is able to organize the flow of internal dynamics into a transient that makes the individual agents more responsive to the subtle changes of the interaction, thereby making it more likely that the ongoing interaction process can be sustained. Here we thus have a concrete example of how *an interaction process can be constitutive of individual behavior*.

It is also worth emphasizing that the processes that drive the necessary internal systemic changes via appropriate input-switching are largely external to the agent. In fact, they are partly constituted by the mutually responsive interaction with the other agent. An agent in an empty 1-D environment would be doomed to linear movement in a single direction, since it is lacking the

ability to internally switch between the two attractor landscapes. Only during an interaction with a responsive partner is the agent's internal organization transformed so as to allow for an open-ended entrainment that can flexibly proceed in either direction.[9] In other words, here we also have a concrete example of how *an interaction process can be constitutive of individual agency*.

3.3 Further investigations of perceptual crossing

To further illustrate the constitutive role of the interaction process on the behavior of the individual agents, Froese and Di Paolo conducted a series of additional experiments with the same computer model which we will briefly describe here. The aim of these models is to give a better sense of how the properties of the interaction process can shape individual behavior.

Receptor field switching experiment

In a first variation of the experimental setup, the receptor fields are switched between the agents such that each agent receives the other's sensory input. This modification cripples the agents' ability to interact with their environment on the basis of coherent sensorimotor correlations created by their own exploratory behavior. Nevertheless, it is found that even under this impaired condition stable perceptual crossing reliably emerges from the inter-agent interactions. Thus, even without any consistent sensorimotor correlations as a basis for individual behavior alone, the inter-individual interaction process essentially negates this lack because of the self-perpetuation of mutually responsive interactions. When the agents interact with each other, the mutuality of the interaction means that they essentially serve as each other's sensor interface, and this mutually and interactively re-established coherence of the individuals' sensorimotor loops reinforces the interaction as a whole.

In this manner, even when most individual behavior is less stable than in the original experimental setup, it is still possible for successful perceptual crossing to self-organize in terms of the relative stabilities of the interaction process. In sum, by modifying the original experimental setup Froese and Di

[9] The crucial role of mutual responsiveness in the scaffolding of individual agency and behavior, as it has been demonstrated by this model, may be able to teach us a lot about how to conduct our social relations. This is true especially in the context of nursing and in other situations of dependency, in particular those involving forms of impaired agency. See Colombetti and Torrance 2009 for a more detailed discussion.

Paolo (2010) thus demonstrated that the interaction process not only makes interaction with the shadow object *unstable*, thereby removing it as a possibility for further entrainment, but that it also plays a constitutive role in making perceptual crossing a *stable* possibility.

Individual behavior versus interaction process experiment

In a second variation, Froese and Di Paolo changed the task so as to introduce a conflict between individual behavior and global stability, namely by further evolving the agents to locate the mobile object which is precisely *not* the other agent, i.e., the other's mobile shadow object. The requirement of detecting social contingency, nevertheless, remains the same as before. This is because the individuals must still distinguish between those interactions that occur with the other's receptor field and those that result from the mobile shadow object, as well as avoid any interaction with the static object. However, in contrast to the original psychological study, here the agents are required to stay with their partner's shadow object, rather than staying with the receptor field of their actual partner. The task is therefore to detect a certain kind of mobile object that gives rise to *non-contingent* interactions, a task that can only be achieved by detecting, and then avoiding, interactions with contingently responsive mobile objects.

It should be noted that, due to the asymmetry inherent in this setup (i.e., agents face in opposite directions, but their shadows are displaced in the same direction), it is impossible for both participants to be interacting with each other's shadow at the same time. Therefore, in order to complete the task it is now necessary for the participants to avoid engaging in inter-individual interaction with each other, so that they can find the shadow object. This will not be easy because (i) engaging in perceptual crossing is still a relatively *stable* behavior, at least for as long as both interactors remain convinced that they are interacting with the other's shadow, and (ii) crossing with the other's shadow remains inherently *unstable*, since that other participant receives no stimulation from this interaction and will therefore keep on looking for the shadow of its partner. In this manner we have created an experimental setup in which the 'intentions' of the individuals and the dynamics of the whole inter-individual interaction process are in direct conflict, and which therefore allows us to further investigate what happens when individuals try to break out of interactions.

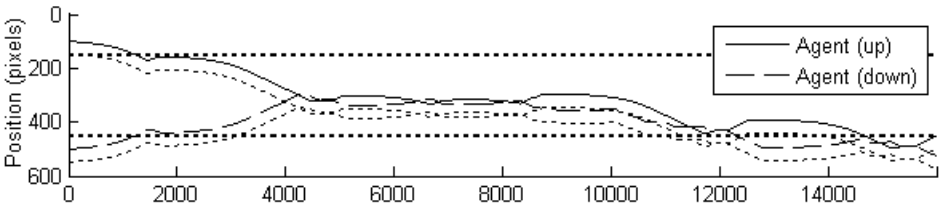


Figure 3: Illustration of the behavior of the agents and their attached shadow objects during a representative trial showing the change in positions over time (figure taken from Froese 2009, p. 164). They first encounter their respective static objects (seen as dotted lines), then continue searching, and finally locate each other and establish perceptual crossing until the end of the run (16000 time steps).

Froese and Di Paolo found that agents can temporarily succeed at this task, but only by regularly falling back into stable patterns of perceptual crossing. The beginning of a representative trial run is depicted in *Figure 3*. At time $t = 0$ the agents begin to move and are briefly distracted by encountering their respective static objects, until they first cross each other's receptor field after $t = 4000$. There then begins an extended period of mutual perceptual crossing until, after $t = 8000$, agent ('down') tries to break out of the interaction with the other agent, and to interact with the other's non-contingent shadow object. This is relatively successful until the agents fall back into mutual perceptual crossing at around $t = 12000$ and again at around $t = 15000$. The modeling results of this experimental variation therefore provide us with a simplified illustrative example of how it can be difficult for individuals to counteract the stability of mutual entrainment in an inter-individual interaction process, even if the disengagement from each other is in fact necessary or beneficial for the completion of their individual tasks.

3.4 Discussion of Results

In earlier sections we suggested that a central claim of the enactive approach to inter-individual interaction is that the dynamics of the interaction process as a whole can play a constitutive role for individual behavior (including sense-making) and for agency. The modeling results presented in the current section of the paper show that this central claim can be systematically approached in a scientific manner without recourse to some social mysterianism. By

constructing minimalist models with very simple artificial 'agents' we have been able to demonstrate how the inter-individual interaction process, *taken as a whole system*, can have important properties that in principle can neither be *separated* from the being and doing of the interacting individuals, nor be *reduced* to the being and doing of those individuals alone.

It may be argued that these modeling results are based on artificial 'agents' that are so minimalist that they have limited value when it comes to a scientific understanding of interactions between human subjects. And, of course, we do not want to claim that these kinds of model agents are actual agents or that the models include all that is essential for proper agency.[10] On the contrary, it further strengthens our arguments if we in fact choose to treat these agents merely as simple dynamical systems, since it shows that we are must be kept in mind that these models are directly based on actual psychological studies, that they replicate the main results of those studies, and that they give additional insight into these results that were not directly evident before (Di Paolo *et al.* 2008). They allow us to distill the essential features of an experimental setup and to explore the space of possible solutions in a mutually informing manner (Froese and Di Paolo in press). For instance, some initial exploratory psychological experiments of perceptual crossing conducted by De Jaegher and Di Paolo at the University of Sussex in 2009 have indicated that human subjects are indeed capable of overcoming the limitations of switched receptor fields because of the stabilizing influence of mutual interaction (Di Paolo, personal communication).

3.5 Future work: Situating interaction processes in a socio-cultural context

We have described and analyzed two types of inter-individual interaction processes in some detail, namely those in which the goals of the interacting individuals are complementary, and those that are in conflict, with the organization of the interaction process itself. But there is another important type of inter-individual interaction which we have not mentioned yet. It can happen that the way in which an interaction process unfolds ends up modifying the goals of the interacting individuals such that there is suddenly a new purpose to their actions. De Jaegher and Di Paolo (2008) discuss an illustrative example where an infant holds up a toy object and in response the

[10] See Froese and Ziemke 2009 for an extended discussion on this topic.

mother also grasps it, and then continues to hold the object when the infant releases its grip. In this situation the infant's action may have started out as a simple stretching of the arm or display and, through the interaction with the mother, the action was in the end invested with a novel social meaning, namely that of giving something to someone else. The behavioral repertoire of the infant has thus been transformed in the interaction, and the act of giving can from now on be initiated intentionally to modulate the flow of social interactions.

Note that this example also nicely illustrates the important distinction between participatory sense-making and social cognition (Gallagher 2009)[11], and a possible transition between the two forms of interaction. As Gallagher points out, while the notion of participatory sense-making denotes the process of sense-making *with* another (although this other is not necessarily the object of this sense-making), social cognition is a term used to characterize the process of cognition *about* another (although this process does not necessarily happen with another). In the case of the mother who spontaneously completes the infant's act of reaching with an object by receiving the object, such that the interaction invests the infant's original act with the new meaning of 'giving', we have an example of how participatory sense-making (the infant-mother interaction provides the infant with new meaning) enables an instance of social cognition (the intentional act of giving involves reference to someone else).

This kind of inter-individual interaction considerably complicates the picture of the constitutive role of the interaction process, because it is no longer just a matter of how an individual's goals are hindered or facilitated by that interaction process. We are now moving toward a more dynamic view of social interaction according to which the interaction process itself can modify the normative structure of the interacting agents while they are interacting. Moreover, this example illustrates how closely the real-time dynamics of an inter-individual interaction process and the historical normative order that defines the socio-cultural background are related. It is on this basis that an individual's enculturation into a social network whose interactions are explicitly and implicitly regulated by an arbitrary (symbolic and traditional) set of preexisting norms becomes a possibility.

We find this kind of inter-enaction of meaning and purpose in an exemplary

[11] Gallagher's use of the term 'social cognition' is in line with the current conventions in the cognitive sciences. But in this context it may be a misleading term for a number of reasons (see next section for details).

form in the case of artistic group activities, such as collective jazz improvisation. In relatively simple versions of joint improvised playing, two or more musicians may agree a basic framework (e.g., chord sequence, tempo) but such a framework is by no means inevitable. One or other player may take the 'lead' while others 'follow' but roles may be swapped rapidly – or perhaps there is no clear lead-follower differentiation. What typically results is a set of unrehearsed and unplanned developments in the musical production which may take off in risky directions that are completely unanticipated by all participants, often radically departing from the set melodic and chordal structure, tempo, and so on. In such explorations there will be a continual and subtle cross-play between what occurs intentionally and what occurs by happenstance, between what is the result of individual agency and what emerges as a group product, and between what is spontaneously co-created in the moment and what is derived from a longstanding heritage of musical tradition. The example of jazz improvisation therefore provides a very useful phenomenon for clarifying how different individual, interactional and socio-cultural factors can shape individual and group behavior.

We currently do not know of any agent-based models which specifically investigate the relationship between the autonomous dynamics of the inter-individual interaction process and the pre-existing normative order that is determined by socio-cultural context in which the interaction process is situated, although there are some promising leads. It may be useful for future Evolutionary Robotics work in this area to take a closer look at some of the models inspired by duet interactions (e.g., Di Paolo 2000, Ikegami and Iizuka 2007) and models of spontaneous goal switching (e.g., Iizuka and Di Paolo 2007b). It is possible that an integration of these two approaches could provide a first step toward a better dynamical understanding of how individual behavior, an ongoing interaction process, and a pre-existing history of interactions together can lead to changes of an agent's goals. At least one thing is clear already: since the enactive approach to agency is going to draw on cognitive science, interaction science, and social science in one unified framework, it is essential to be clear about the term 'social', which perhaps has different connotations in each of these areas of research.

4. Inter-Enaction and 'Sociality'

In the preceding discussion we have sometimes referred to 'interaction' (or 'inter-enaction', our preferred term for a certain enactive view of the latter), and sometimes to 'social interaction'. We act and cognize with others and about others in our world in a variety of ways – are all of those ways to be included in a blanket way within the terms of the account of 'participatory sense-making' (PSM) sketched here and in other works cited? In what follows we will consider a view which is, in many ways, critical of the PSM or inter-enactive account. This discussion will enable us to clarify what we see as a correct evaluation of the scope of PSM theory, and also to put right certain possible mistaken assumptions about the PSM approach.

In a recent paper, Pierre Steiner and John Stewart (2009) claim that the PSM account put forward by De Jaegher and Di Paolo, and endorsed by others (including ourselves) is, at worst, radically flawed, and at best, much more limited in the extent of its application than its proponents are claiming.[12] This is for two interconnected reasons. First, the PSM view fails to take account of the fundamental role played by social norms, or (as they put it) 'normative order', in setting the context for our inter-individual interactions. These social norms include communicative, moral, legal, economic, religious, etc. rules, expectations, forms of life, and so on. As they see it, these normative structures constitute the very fabric of the *social* environment in which humans live and interact on a day-to-day basis.

Second, far from being a field of autonomy, the realm of social normativity imposes important constraints (they claim) on how the fine-grained interactions of our day-to-day life unfold: the existence and ubiquity of such constraints make it appropriate to talk of this field of inter-individual interactions as one of *heteronomy*, rather than – as the PSM account suggests – one of *autonomy*.

The authors argue that the notion of 'sociality' – and related terms such as 'social cognition', 'social interaction' – can be understood in at least two importantly different ways. On their own view, which makes strong appeals to a tradition of social theory stemming from Émile Durkheim, Talcott Parsons and many others, sociality is largely constituted by this pre-existing, culturally inherited, normative order that each social agent (human) finds him/herself

[12] Their account is couched in terms of a discussion of 'social cognition' but it equally applies to 'social interaction', and indeed to the nature of the 'social' in general.

embedded in throughout daily life. The norms in this pre-existing structure

> actually constitute the possibility of enacting worlds that would just not exist without them. Interactions between two or more agents are never properly social unless they take place in the context of an environment of social structures or norms which give meaning to the interactions. (Steiner and Stewart 2009, p. 528)

Let us call Steiner and Stewart's account the 'social normative order' approach (SNO for short).

On the contrasting view of sociality, which they identify with the PSM account, sociality emerges from the dynamics of the inter-individual interactions as they unfold in the here-and-now. The relatively small-scale interactions that are the major focus of the PSM account are actually 'heteronomous' with respect to these large-scale pre-existing structures, rather than autonomous processes that are constitutive of sociality.

We believe that the SNO account makes some crucially important points, and does indeed highlight gaps or inadequacies in the original PSM account. Nevertheless we will argue that it is possible to resolve the apparent disparities between the PSM and SNO accounts, by making some conceptual clarifications, and by delineating different terms of reference or scopes of application for the different accounts. The result will be a fuller picture of interactive agency and of sociality than is presented in either account as they stand.

Consider again the situation where two people are walking towards each other along a confined passageway. The PSM account will refer to this situation as involving an independent dynamic of interaction, which has its own *autonomy,* which in turn constrains the activities of the individual participants in the situation. Yet clearly, to the extent to which the individual participants in such a corridor scene are 'subject to' this dynamic, they are 'heteronomous' with respect to the dynamic itself. So heteronomy could be seen as being equally a feature of the PSM account as provided by De Jaegher and Di Paolo (although their account happens not to employ that term). Of course any actual corridor scene will include a host of other features which are not specified in the bare description 'two people rapidly walking towards each other and attempting to adjust their position within the narrow space of the passageway' – where it is more or less treated as a physical interaction. There will be rules of etiquette, for example, that prescribe ways of dealing with the situation that are and aren't 'socially acceptable' (one person shouting at or shoving past, the

other will be considered rude or even an assault; laying on the ground and inviting the other to walk over you would be considered impossibly obsequious; and so on).[13] These will be part of a vast array of culturally prescribed social norms – sometimes explicitly codified, and sometimes implicitly understood and even unconscious to participants – that govern the way that people are expected or allowed to behave in public spaces. These do indeed shape the situation that often unfolds in the way described in the canonical 'corridor scene'. And indeed, in relation to these pre-existing normative structures, the dynamic of the interaction of the two passing figures, as they alternately move to one side, then to the other, of the passageway, will indeed, qua interactive dynamics, be describable as heteronomous rather than autonomous. But at the same time, in relation to the individual participants themselves, this interactive process does occupy a degree of autonomy, as described in the PSM account.

For a similar reason, the pre-existing normative orders that are referred to in the SNO account can perhaps equally be described as 'autonomous' (independent; transcendent) with respect to the participants. Steiner and Stewart choose not to use the word 'autonomous' in this connection (2009, p. 530) because they want to stress the idea of heteronomy (when focusing on the participants). But of course 'autonomy' and 'heteronomy' are (certainly in this context) point-of-view-relative terms. In the case of both PSM and SNO there are individual agents and a supra-individual structure. In each case the supra-individual structure (interaction-dynamic in the PSM account; historically-given normative order in the SNO account) has 'autonomy' in the broad sense of being 'independent', having its own 'life'. Also, in both accounts this structure constrains and enables actions by the individuals. Conversely, in the case of both accounts, the individual agents are heteronomous with respect to the over-arching structure, because of the constitutive, enabling role each structure has on their activity.

Once pointed out, this should seem obvious, but perhaps it needs to be clarified, so as to forestall any further confusion. Thus Steiner and Stewart are perhaps wrong in saying that heteronomy plays no role in the PSM account (because it is there, even though not named as such). Nevertheless they are correct in saying that the kind of heteronomy imposed on (or better, implied

[13] See Colombetti and Torrance 2009, for an elaboration of this point, and for a discussion of the richly *ethical* nature of such apparently simple interactions.

for) individual agents by the normative order which provides the medium for their interactions (and makes them 'fully human agents') has complex, wide-ranging and subtle characteristics, which are not recognized (or not stressed fully enough) in the PSM account, and which are elaborated at length within Steiner and Stewart's paper. Conversely, the PSM account in turn involves subtle features – for example those to do with the dynamics of interaction which we have sought to stress in the descriptions of experimental investigations earlier in the paper – which are not taken up in Steiner and Stewart's SNO account. Thus each paper contains important elements that have to be brought together in order to have a properly filled-out picture of social inter-(en)action.

Another, related, point that needs to be made concerns the 'sense-making' aspect of the PSM account. In the above discussion we have mostly concentrated on the autonomy of the interaction between participants, relative to those individual agents themselves. But of course, as the term 'participatory sense-making' is intended to convey, the interaction between two or more agents (in the face-to-face, real-time situations which were of primary interest to the authors of the PSM account) typically involves a continual exploratory unfolding of the situation.

Consider, as an example, the interaction that might typically occur between two motorists who find themselves in a collision: as they encounter one another each may have a pre-planned culturally determined 'script' which they may aspire to follow in a way that will (they hope) remain relatively impervious to the way the other may seek to influence the interaction. Yet what often happens is that the actual development of the interaction involves a path which is mutually influenced by the two actors, and which often follows a trajectory that conforms to the prior expectations of neither of them.

Talking of the 'autonomy' of the interaction process helps to evoke the way that this trajectory seems to take on a 'life' of its own, to a greater or lesser degree independent of the individual participants. But equally, one can talk of this unfolding as a mutual exploration of the relation-space, where significances are jointly created (indeed, 'enacted') by the participants. For example there will be a negotiation over the affective tone that this encounter will take – will the course it follows be on the whole friendly or hostile? Sometimes such joint meaning-making will be primarily cooperative or collaborative, sometimes it will have a primarily combatitive or aggressive character; more often than not it will have elements of both.

There are many other examples of encounters that facilitate an ongoing exploration that has an exploratory, creative character on the part of the actors, where meanings are constantly 'enacted', 'challenged', 'reinforced', and so on but where this exploratory, enactive process may also be seen as having its own independent dynamic. The example of jazz improvisation was mentioned earlier. (Even in this sphere of artistic collaboration there may be both cooperative and confrontational elements). To get a feel for other kinds of example think of the kinds of interactions that commonly take place between people, whether in buses or underground trains or lifts, etc., people playing sports like football and squash; or again people in various kinds of sexual encounters, whether they be courtship scenes, blocked or reluctantly-borne come-ons, full-blooded passion, or any of the other myriad variants of sexual interaction.

How, then, should we relate these two accounts or these two levels of description? How does the exploratory, enactive, and immediate character of the meaning-making dynamic in real-time interactions, as characterized within the PSM story, cohere with the vast edifice of inherited, culturally-accreted norms which is the dominant *motif* of the SNO account? There may seem to be a conflict: surely, it might be said, both cannot be true. Yet that is just what, on a more reflective examination, can be agreed to be indeed the case. We construct the shared meanings in our ongoing, real-time interactions, within the context of a vast array of social 'givens', which have a solidity for individual participants – a social solidity, one might say. These social 'givens' (both informal and codified) will both facilitate and constrain the individual interactive encounters that occur at the face-to-face level. Thus, for the disputants in the automobile collision these norms include legal regulations, financial constraints, bounds of moral acceptability in word and action, instructions or recommendations in documents on how to conduct oneself at an accident scene, as well, of course, as the physical and technological conditions of the situation itself and the perceptions and memories of the sequence of relevant events.[14] But while these social givens set prescriptive

[14] Of course there are many other features of social encounters besides the interactional dynamics highlighted by PSM and the historic normative structures highlighted by SNO. It should be obvious that physical, biological, psychological conditions of different sorts play important roles, of both a primary or supportive kind. The roles of different kinds of artefacts, including texts and other symbolic media, and technological devices of many kinds, should also be stressed (e.g., Clark 2003). PSM and SNO mark out important necessary features of social action: they are in no way to be considered as jointly sufficient.

and/or permissive conditions for the interactions that occur on a given occasion, the interactions themselves will involve creative reinterpretation and modification of the very norms which are the framework within which the interaction takes place. As often as not these reinterpretations are trivial. Sometimes they can be of major cultural or political significance – as, for example, was the occasion on December 1st 1955 in Montgomery Alabama when Mrs. Rosa Parks, a black passenger on a bus, refused to move from a her seat to enable a white person entering the bus to sit in a whites-only row, in accordance with racial segregation practices in operation at the time.

Moreover, it is worth asking how these apparently impersonal social norms that are emphasized by Steiner and Stewart actually maintain their continued existence. They don't just exist in a special normative realm independently of the actual lives of people: they are embedded in the ways people conduct those lives – their continued existence requires that they be continually (inter-) enacted, in either word or deed. As pointed out above, more often than not norms are written down in various forms (or are repeated in various kinds of confirmatory speech-act). But this is true only of some kinds of norms, and even those will actively maintain their force in the social order only as a result of compliant patterns of action and interaction, and through acts of positive and negative sanction. Thus what made the whites-only norm a norm that was in force in buses in the Alabama of 1955 was the fact that it was regularly adhered to in action by both white and black travelers, and that non-compliance was met with fines or other punishments. So, while Steiner and Stewart are right that interactions of the face-to-face sort take on the character that they do because of constitutive role played by the background of historic social norms, those historic norms themselves are perpetuated through continuing compliant interactions by the members of the population for whom those norms have force.[15]

What these considerations strongly suggest is that the PSM account and the

[15] Indeed it is important to see that the term 'participatory sense-making' should be interpreted to cover, not just the kind of case where new interpretative directions are taken for a given rule or set of rules, but also the kind of case where an existing way of doing things is reaffirmed by faithful repetition. Thus taking a moment to say Grace before starting a family meal will be as much a case of participatory sense-making, as breaking with tradition by missing out on Grace. In the latter case the participants are creating a new 'sense' to their joint meal-taking activity; in the former case the participants are re-affirming their recognition of an old 'sense' – that of their traditional way of starting a meal-time. In each case the participants are *enacting* a continuation into the future of what they perceive as the way the past demands or permits them to act in the here and now.

SNO accounts are both necessary to a full understanding of inter-individual relationships and larger scale social relationships as well, of course, of individual agency. Not merely are they both necessary, but they are complementary processes, in that each process is partially constitutive of the other process. As Steiner and Stewart have emphasized in their account, the interactions that take place in real time, at a face-to-face level are constitutively governed by countless historically accreted social norms that exist as an impersonal background to the real-time interactions. However, as we have argued, the PSM account of face-to-face interactions[16] gives an account of the social reality of those social norms, by explaining that the existence of the historic force of those social norms is itself constituted by countless interactions, sayings and collaborations in the past; and that their continued existence is constituted by further interactions, sayings and collaborations into the future.

5. Conclusions

In this paper we have discussed three kinds of view: individualism; PSM; SNO. These three views suggest three different levels of analysis: one focusing on individual 'in-the-head' cognitive processes; one on the dynamics of inter-individual interactions; and on the historical structures of large- (and small-) scale social norms. Individualism interprets the inter-individual and the social in terms of individual acts and internal processes, more often than not couched in terms of some variant of the cognitivist (sense-model-plan-act) story, whose odd solipsist character only becomes evident when critics of the story, such as enactivists and others, bring it to attention. The enactive approach to agency, with its emphasis on the relational nature of life and mind, provides a different kind of departure point for a consideration of the role of sociality.

We have considered two variants here. PSM concentrates on the inter-individual level, the level at which people participate with each other in a shared moving present, and a shared presence, in which the dynamic of the interaction can be seen as having its own relative autonomy, both arising out of the agents' moves and as continually restructuring them. We have shown how

[16] The term 'face-to-face' should be used with some caution. Some interactions, for example via Facebook or Twitter, are hardly face-to-face in any literal sense. Nor are they necessary small-scale or intimate: a given announcement on a social networking site may have an audience of millions!

theoretical claims about the dynamics of these interactions can be grounded in experimental models based on minimalist scenarios using artificial agents. SNO sees action primarily in terms of the historic social norms which have created the background of expectations and rules that act as the fundamental enabling and constraining factors of the unfolding shared present. PSM and SNO can be seen as offering two contrasting responses to classical individualism, each of which stresses a crucial aspect of the supra-individual nature of human action.

However these two views are not in competition. Clearly they offer necessary complements to each other. PSM needs SNO to explain the sense in which present-tense interactions are truly social, rather than just 'inter-agential'. But SNO needs PSM to explain how the vast edifice of historical normativity left by dead people and dead time, retains its liveness in the present and into the future by countless collaborative acts of reinterpretation, revision and reaffirmation. A considered version of inter-enactivism has to stress both these levels as offering important constitutive conditions for human action.

Thus it is important, as social normative order theory insists, to see every action as taking place within a historical context – which includes the high-level accreted norms at various scales of globality and locality – broadly universal rules to do with economy, morality, prevailing technical conditions, etc., as well as community-specific and family-specific normative environments; but, it must be stressed, much else besides social norms: the physical and biological conditions, the phylogenetic and ontogenetic inheritances, and so on. However, at the moment of action or interaction itself there is also the dynamic of how the actors in a situation both are shaped by these normative conditions and reshape them in their interaction, and how the actions of each individual agent in the situation both shape and are shaped by the actions of the others present in that situation. This is the domain of participatory sense-making, but for a more complete enactivist picture we need to combine this domain of inter-individual dynamic presence with the past social conditions which have brought those individuals to this presence.

REFERENCES

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology, 27*(1), 32-47.

Barandiaran, X., Di Paolo, E. A., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior, 17*(5), 367-386.

Barandiaran, X., & Ruiz-Mirazo, K. (Eds.) (2008). Special Issue on "Modeling Autonomy". *BioSystems Journal, 91*(2).

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. J. Mataric, J.-A. Arcady, J. Pollack & S. W. Wilson (Eds.), *From Animals to Animats 4: Proc. of the 4th Int. Conf. on Simulation of Adaptive Behavior*, (pp. 421-429). Cambridge, MA: MIT Press.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior, 11*(4), 209-243.

Boden, M. A. (2006). Of islands and interactions. *Journal of Consciousness Studies, 13*(5), 53-63.

Brooks, R. A. (1991). Intelligence without reason. In J. Myopoulos & R. Reiter (Eds.), *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence*, (pp. 569-595). San Mateo, CA: Morgan Kaufmann.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford: Oxford University Press.

Colombetti, G., & Torrance, S. (2009). Emotion and ethics: An inter-(en)active approach. *Phenomenology and the Cognitive Sciences, 8*(4), 505-526.

De Jaegher, H., & Di Paolo, E. A. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences, 6*(4), 485-507.

De Jaegher, H., & Di Paolo, E. A. (2008). Making sense in participation: An enactive approach to social cognition. In F. Morganti, A. Carassa & G. Riva (Eds.), *Enacting Intersubjectivity: A Cognitive and Social*

Perspective on the Study of Interactions, (pp. 33-47). Amsterdam: IOS Press.

De Jaegher, H., & Froese, T. (2009). On the role of social interaction in individual agency. *Adaptive Behavior, 17*(5), 444-460.

Di Paolo, E. A. (2000). Behavioral coordination, structural congruence and entrainment in a simulation of acoustically coupled agents. *Adaptive Behavior, 8*(1), 25-46.

Di Paolo, E.A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429-452.

Di Paolo, E.A. (Ed.) (2009). Special issue on "The Social and Enactive Mind". *Phenomenology and the Cognitive Sciences, 8*(4).

Di Paolo, E. A., Rohde, M., & De Jaegher, H. (2010). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart, O. Gapenne & E. A. Di Paolo (Eds.), *Enaction: Towards a New Paradigm for Cognitive Science*, (pp. 33-87). Cambridge, MA: MIT Press.

Di Paolo, E. A., Rohde, M., & Iizuka, H. (2008). Sensitivity to social contingency or stability of interaction? Modelling the dynamics of perceptual crossing. *New Ideas in Psychology, 26*(2), 278-294.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences, 3*(1), 63-73.

Froese, T. (2009). *Sociality and the Life-Mind Continuity Thesis: A Study in Evolutionary Robotics*. D.Phil. thesis, Brighton, UK: University of Sussex.

Froese, T., & Di Paolo, E. A. (2008). Stability of coordination requires mutuality of interaction in a model of embodied agents. In M. Asada, J. C. T. Hallam, J.-A. Meyer & J. Tani (Eds.), *From Animals to Animats 10: Proc. of the 10th Int. Conf. on Simulation of Adaptive Behavior*, (pp. 52-61). Berlin, Germany: Springer-Verlag.

Froese, T., & Di Paolo, E. A. (2009). Sociality and the life-mind continuity thesis. *Phenomenology and the Cognitive Sciences, 8*(4), 439-463.

Froese, T., & Di Paolo, E. A. (2010). Modeling social interaction as

perceptual crossing: An investigation into the dynamics of the interaction process. *Connection Science, 22*(1), 43-68.

Froese, T., & Di Paolo, E. A. (in press). Toward minimally social behavior: Social psychology meets evolutionary robotics. *Advances in Artificial Life: Proc. of the 10th Euro. Conf. on Artificial Life.* Berlin: Springer-Verlag.

Froese, T., & Gallagher, S. (2010). Phenomenology and artificial life: Toward a technological supplementation of phenomenological methodology. *Husserl Studies, 26*(2), 83-106.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence, 173*(3-4), 366-500.

Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies, 8*(5-7), 83-108.

Gallagher, S. (2005). *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17*(2), 535-543.

Gallagher, S. (2009). Two problems of intersubjectivity. *Journal of Consciousness Studies, 16*(6-8), 289-308.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. P. Racine, C. Sinha & E. Itkonen (Eds), *The Shared Mind: Perspectives on Intersubjectivity*, (pp. 17-38). Amsterdam: John Benjamins.

Gallagher, S., & Zahavi, D. (2008). *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science.* London: Routledge.

Harvey, I., Di Paolo, E. A., Wood, R., Quinn, M., & Tuci, E. A. (2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial Life, 11*(1-2), 79-98.

Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind and Language, 19*(5), 548-573.

Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. In D. Hutto (Ed.), *Narrative and Understanding Persons*, (pp. 43-68). Cambridge: Cambridge University Press.

Iizuka, H., & Di Paolo, E. A. (2007a). Minimal Agency Detection of Embodied Agents. In F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey & A. Coutinho (Eds.), *Advances in Artificial Life: Proc. of the 9th Euro. Conf. on Artificial Life*, (pp. 485-494). Berlin, Germany: Springer-Verlag.

Iizuka, H., & Di Paolo, E. A. (2007b). Toward Spinozist robotics: Exploring the minimal dynamics of behavioral preference. *Adaptive Behavior, 15*(4), 359-376.

Ikegami, T., & Iizuka, H. (2007). Turn-taking interaction as a cooperative and co-creative process. *Infant Behavior & Development, 30*(2), 278-288.

Kiverstein, J., & Clark, A. (2009). Introduction. Mind embodied, embedded, enacted: One Church or Many? *Topoi, 28*(1), 1-7.

Maturana, H. R., & Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambhala Publications.

Menary, R. (Ed.) (2010). Special issue on '4E Cognition: Embodied, Embedded, Enacted, Extended'. *Phenomenology and the Cognitive Sciences, 9*(4).

Merleau-Ponty, M. (1962). *Phenomenology of perception*. (Tr. by C. Smith). New York: Routledge & Kegan Paul. [1945]

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Nolfi, S., & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. Cambridge, MA: MIT Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 939-1031.

Port, R. F., & van Gelder, T. (Eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Steiner, P., & Stewart, J. (2009). From autonomy to heteronomy (and back): The enaction of social life. *Phenomenology and the Cognitive Sciences, 8*(4), 527-550.

Stewart, J., Gapenne, O., & Di Paolo, E.A. (Eds.) (2010). *Enaction: Towards a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press.

Thompson, E. (Ed.) (2001). *Between Ourselves: Second-Person Issues in the Study of Consciousness*. Thorverton, UK: Imprint Academic. Also published as a special issue of the *Journal of Consciousness Studies, 8*(5-7), 1-309.

Thompson, E. (2005). Sensorimotor subjectivity and the enactive approach to experience. *Phenomenology and the Cognitive Sciences, 4*(4), 407-427.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: The Belknap Press of Harvard University Press.

Torrance, S. (2002). The skill of seeing: Beyond the sensorimotor account. *Trends in Cognitive Sciences, 6*(12), 495-496.

Torrance, S. (2005). In search of the enactive: Introduction to special issue on enactive experience. *Phenomenology and the Cognitive Sciences, 4*(4), 357-368.

Torrance, S. (Ed.) (2005). Special issue on Enactive Experience, 1. *Phenomenology and the Cognitive Sciences, 4*(4).

Torrance, S. (Ed.) (2007). Special issue on Enactive Experience, 2. *Phenomenology and the Cognitive Sciences, 6*(4).

Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition, 34*(1), 72-87.

Varela, F. J, Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences, 1*(2), 97-125.

Strong Interaction and Self-Agency

Shaun Gallagher *
gallaghr@mail.ucf.edu

ABSTRACT

The interaction theory of social cognition contends that intersubjective interaction is characterized by both immersion and irreducibility. This motivates a question about autonomy and self-agency: If I am always caught up in processes of interaction, and interaction always goes beyond me and my ultimate control, is there any room for self-agency? I outline an answer to this question that points to the importance of communicative and narrative practices.

In regard to social cognition, there has been growing opposition to the standard theory-of-mind (ToM) views, usually referred to as theory theory (TT) and simulation theory (ST). I have defended an alternative approach: "interaction theory" (IT). IT is based on evidence from both phenomenology and developmental psychology, and it offers an alternative to the simulation interpretation of the neuroscience of mirror neurons. An important part of IT is its emphasis on 'strong interaction' (Gallagher in press; also see De Jaeger *et al.* 2010) – a concept of interaction that is a seemingly pervasive feature of intersubjectivity. In this paper I take a closer look at this concept and raise questions about what appears to be a threat to the notion of self-agency. The question is: If we are so interactively interdependent on others in our everyday practical and communicative behaviors, is there any room for autonomy?

Interaction Theory (IT) as an Alternative to TT and ST

In psychology, philosophy of mind, and more recently, in the neurosciences, studies of how one person understands and interrelates with another person

* Department of Philosophy, University of Central Florida – Institute of Simulation and Training – University of Hertfordshire – University of Copenhagen

have been dominated by two main approaches: theory theory and simulation theory. The major tenets of TT are based on scientific experiments that show that children develop an understanding of other minds around the age of four. One version of TT claims that this understanding is based on an innately specified, domain specific mechanism designed for 'reading' other minds (e.g., Baron-Cohen 1995, Leslie 1991). An alternative version claims that the child attains this ability through a course of development in which the child tests and learns from the social environment (e.g., Gopnik and Meltzoff 1997). Common to both versions is the idea that children attain their understanding of other minds through the use of folk or commonsense psychology which they use to make theoretical inferences about certain entities to which they have no access, namely, the mental states of other people. When we make such inferences and attribute specific mental states to others, we are said to be *mentalizing* or *mindreading*.

ST, in contrast, argues that rather than theorizing or making inferences about the other person's mind, we use our own mental experience as an internal model for the other mind (e.g., Gordon 1986, 1995; Heal 1986, 1998a, 1998b). To understand the other person, I simulate the thoughts or feelings that I would experience *if I were in the situation of the other*, exploiting my own motivational and emotional resources. I imagine what must be going on in the other person's mind; or I create in my own mind pretend beliefs, desires or strategies that I use to understand the other's behavior. My source for these simulations is not a theory that I have. Rather, I have a real model of the mind at my immediate disposal, that is, I have *my own mind*, and I can use it to generate and run simulations. I simply run through the sequence or pattern of behavior or the decision-making process that I would engage in if I were faced with the situation in question. I do it 'off line', however. That is, my imaginary rehearsal does not lead to actualizing the behavior on my part. Finally, I attribute this pattern to the other person who is actually in that situation.

Despite extensive debates between proponents of TT and ST, respectively, TT and ST share three basic suppositions. The three suppositions are these.

(1) The problem of social cognition is due to the lack of access that we have to the other person's mental states. Since we cannot directly perceive the other's thoughts, feelings, or intentions, we need some extra-perceptual cognitive process (mindreading or mentalizing) that will allow us to infer or simulate what they are.

(2) Our normal everyday stance toward the other person is a third-person, observational stance. Based on what we observe we use mindreading to *explain* or *predict* their behaviors.

(3) These mentalizing processes constitute our primary and pervasive way of understanding others.

There are also a number of unsolved problems associated with these ToM approaches. I won't go into detail here, but I'll give a brief indication of some of these problems.[1] First, some (but not all) theorists claim the process of theoretical inference or simulation is conscious or introspective (e.g., Goldman 1995; Goldman 2006, p. 147); but there is no phenomenological evidence that this is so, and there should be if the process is both consciously explicit and pervasive. That is, we should be able to catch ourselves in the act, but we don't. The second problem is what I refer to as the starting problem, a version of the frame problem. For TT, the question is how do I know what piece of folk psychology (what rule, or what platitude) actually applies to the case at hand. For ST, one can see the problem clearly in the following description of a simulation routine provided by Nichols and Stich:

> The basic idea of what we call the 'off-line simulation theory' is that in predicting and explaining people's behavior we take our own decision making system 'off-line', supply it with 'pretend' inputs that have the same content as the beliefs and desires of the person whose behavior we're concerned with, and let it make a decision on what to do. (Nichols and Stich 2003, pp. 39-40)

Simulation as a form of mindreading is supposed to provide insight into the beliefs and desires of the other person, but it seems that we need to know the content of those mental states in order to do the simulation. Neither TT nor ST provide a good answer to the starting problem.

A third problem concerns diversity and applies specifically to ST. Keysers and Gazzola describe simulation in the following way:

> In [simulation] cases, observing what other people do or feel is therefore transformed into an inner representation of what we would do or feel in a similar situation – as if we would be in the skin of the person we observe. (Keysers and Gazzola 2006, p. 390)

But how does knowing what we would do help us know what someone else

[1] See Gallagher 2005 and 2007 for more detail.

would do? Indeed, many times we are in a situation where we see what someone is doing, and know that we would do it differently, or perhaps, not do it at all. A fourth problem concerns development. The kind of inferential or simulation processes found in explicit versions of TT and ST are too cognitively complex to account for the infant's ability to understand the intentions of others. Yet, as we'll see below, there is a large amount of evidence to support the idea that young infants are able to grasp the intentions of others.

The developmental problem is addressed by a recent version of ST that relies on an interpretation of mirror neuron (MN) activation as a form of simulation. In this case, simulation is said to be fast and automatic. If MNs are active in young infants, then the developmental problem does not apply to this version of ST. Since activation of MNs are non-conscious, the issue of phenomenological evidence is irrelevant, and there is no starting problem. So-called neuronal ST, then solves all of the above problems except perhaps the diversity problem. But there are other problems involved in neural ST. One concerns the fact that simulation is originally defined as involving pretense. As Nichols and Stich make clear in the above quote, simulation involves the use of pretend beliefs and desires. We pretend to be in the other person's shoes in order to run the routine. But the notion of pretense does not apply in the case of MNs. Indeed, most theorists claim that MNs are neutral with respect to who the agent is, and agent-neutrality is not consistent with the notion of pretense. MNs can't account for me pretending to be you if in fact there is no distinction between me and you at that level. As a result, there have been attempts to shift the definition of simulation to involve a simple matching (e.g., Goldman 2006, Rizzolatti *et al.* 2001). My motor system is said to go into a state matching yours when I see you perform an action. But the neurological details do not bear this out[2], and it seems counter-intuitive if we think of how we interact with others. In the majority of cases we are not imitating or mimicking others; rather, our motor systems are busy supporting responses or complementary actions.

This is not an exhaustive list of problems with TT and ST, but it should be sufficient to see why we might want to find a better account of social cognition. Interaction theory is proposed as that better account. IT challenges the three suppositions associated with ToM approaches. In their place IT argues for the following propositions.

[2] See, e.g., Catmur *et al.* 2007, Dinstein *et al.* 2008.

(1) Other minds are not hidden away and inaccessible. The other person's intentions, emotions, and dispositions are expressed in their embodied behavior. In most cases of everyday interaction no inference or projection to mental states beyond those expressions and behaviors is necessary.

(2) Our normal everyday stance toward the other person is not third-person, detached observation; it is second-person interaction. We are not primarily spectators or observers of other people's actions; for the most part we are interacting with them in some communicative action, on some project, in some pre-defined relation; or we are treating them as potential interactors.

(3) Our primary and pervasive way of understanding others does not involve mentalizing or mindreading; in fact, these are rare and specialized abilities that we develop only on the basis of a more embodied engagement with others.

IT emphasizes the role of communicative and narrative practices, and it appeals to evidence from developmental studies, starting with primary and secondary intersubjectivity (Trevarthen 1979; Trevarthen and Hubley 1978).

- Primary intersubjectivity (starting from birth) – Sensory-motor abilities – enactive perceptual capacities in processes of interaction
- Secondary intersubjectivity (starting around 1 year of age) – joint attention, shared contexts, pragmatic engagements, acting *with* others
- Communicative and narrative competencies (starting from 2-4 years) – communicative and narrative practices that represent intersubjective interactions, motives, and reasons and provide a more nuanced and sophisticated social understanding.

In this paper I will begin with a focus on primary and secondary intersubjectivity, but I'll return the issue of narrative competence. I take this strategy because I first want to focus on the nature of interaction itself, and most of the essential aspects can be grasped in primary and secondary intersubjectivity. When it comes to the question of self-agency, however, narrative will be shown to play an important role.

The Development of Interaction

Primary intersubjectivity consists of the innate or early-developing sensory-motor capacities that bring us into relation with others and allow us to interact with them. These capacities are manifested at the level of perceptual experience -- we *see* or more generally *perceive* in the other person's bodily movements, gestures, facial expressions, eye direction, etc. what they intend and what they feel, and we respond with our own bodily movements, gestures, facial expressions, gaze, etc. From birth the infant is pulled into these interactive processes. This can be seen in the very early behavior of the newborn. Infants from birth are capable of perceiving and imitating facial gestures presented by another (Meltzoff and Moore 1977, 1994). Importantly, this kind of imitation is not an automatic or mechanical procedure; Csibra and Gergely (2009) have shown, for example, that the infant is more likely to imitate only if the other person is attending to it.

Primary intersubjectivity can be specified in more detail as the infant develops. At 2 months, for example, infants are able to follow the gaze of the other person, to see that the other person is looking in a certain direction, and to sense what the other person sees (which is sometimes the infant herself), in a way that throws the intention of the other person into relief (Baron-Cohen 1995; Maurer and Barrera 1981). In addition, second-person *interaction* is evidenced by the timing and emotional response of infants' behavior. Infants «vocalize and gesture in a way that seems [affectively and temporally] 'tuned' to the vocalizations and gestures of the other person» (Gopnik and Meltzoff 1997, p. 131). At 5-7 months, infants are able to detect correspondences between visual and auditory information that specify the expression of emotions (Walker 1982; Hobson 1993, 2002). At 6 months infants start to perceive grasping as goal directed, and at 10-11 months infants are able to parse some kinds of continuous action according to intentional boundaries (Baldwin and Baird 2001; Baird and Baldwin 2001; Woodward and Sommerville 2000). They start to perceive various movements of the head, the mouth, the hands, and more general body movements as meaningful, goal-directed movements (Senju *et al.* 2006).

Developmental studies show the very early appearance of, and the importance of, interactive attunement in the form of timing and coordination in the intersubjective context. In still face experiments, for example, infants are engaged in a normal face-to-face interaction with an adult for 1 to 2 minutes,

followed by the adult assuming a neutral facial expression. This is followed by another normal face-to-face interaction. Infants between 3 and 6 months become visibly discouraged and upset during the still face period (Tronick 2007, Tronick *et al.* 1978). The importance of interactive touch has also been demonstrated in the still-person effect (Muir 2002).

Murray and Trevarthen (1985) have also shown the importance of the mother's live interaction with 2-month old infants in their double TV monitor experiment where mother and infant interact by means of a live television link. The infants engage in lively interaction in this situation. When presented with a recorded replay of their mother's previous actions, however, they quickly disengage and become distracted and upset. These results have been replicated, eliminating alternative explanations such as infants' fatigue or memory problems (Nadel *et al.* 1999, Stormark and Braarud 2004).

Primary intersubjectivity is not something that disappears after the first year of life. It is not a stage that we leave behind, and it is not, as Greg Currie suggests, a set of precursor states «that underpin early intersubjective understanding, and *make way* for the development of later theorizing or simulation» (Currie 2008, p. 212, *my emphasis*).[3] Rather, citing both behavioral and phenomenological evidence, IT argues that we don't leave primary intersubjectivity behind; the processes involved here don't "make way" for the purportedly more sophisticated mindreading processes – these embodied interactive processes continue to characterize our everyday encounters even as adults. That is, we continue to understand others in strong interactional terms, facilitated by our recognition of facial expressions, gestures, postures, and actions as meaningful.

Scientific experiments bear this out. Point-light experiments (actors in the dark wearing point lights on their joints, presenting abstract outlines of emotional and action postures), for example, show that not only children (although not autistic children) but also adults perceive emotion even in movement that offers minimal information (Hobson and Lee 1999, Dittrich *et al.* 1996). Close analysis of facial expression, gesture and action in everyday contexts shows that as adults we continue to rely on embodied interactive abilities to understand the intentions and actions of others and to accomplish interactive tasks (Lindblom 2007, Lindblom and Ziemke 2007).

[3] Cf. Baron-Cohen 1991 and 1995.

By the end of the first year of life, infants have a non-mentalizing, perceptually-based, embodied and pragmatic understanding of the intentions and dispositions of other persons. With the advent of joint attention (at around 9 months) and secondary intersubjectivity (at around 1 year) infants start to use context and enter into situations of participatory sense-making (De Jaegher and Di Paolo 2007). That is, infants begin to co-constitute the meaning of the world in their interactions with others. We start to understand the world through our interactions with others, and we gain a more nuanced understanding of others by situating their actions in contexts that are defined by both pragmatic tasks and cultural practices.

Meaning and emotional significance is co-constituted in the interaction – not in the private confines of one or the other's head. The analyses of social interactions in shared activities, in working together, in communicative practices, and so on, show that agents unconsciously coordinate their movements, gestures, and speech acts (Issartel *et al.* 2007, Kendon 1990, Lindblom 2007). In the contextualized practices of secondary intersubjectivity timing and emotional attunement continue to be important as we coordinate our perception-action sequences; our movements are coupled with changes in velocity, direction and intonation of the movements and utterances of the speaker.

The kind of embodied and contextualized interaction that we find in primary and secondary intersubjectivity is what I am calling 'strong interaction'. In strong interaction, our movements are often synchronized in resonance with others, following either in-phase or phase-delayed behaviour, and in rhythmic co-variation of gestures, facial or vocal expressions (Fuchs and De Jaegher 2009). This kind of intersubjective interaction involves coordination but does not imply perfect synchronization. Non-autistic infants from 3-months of age prefer slight modulations (time-delays) and imperfect contingency in responses (Gergely 2001). As De Jaegher (2008) suggests, continuous movements between synchronised, desynchronised and the states in-between, drive the process. Attunement, loss of attunement, and re-established attunement maintain both differentiation and connection.

A Closer Look at Interaction

I want to focus on two aspects of strong interaction: immersion and irreducibility. The first involves the idea that interaction is not something that we decide to enter into. Rather, it is, as the existentialists might say, something that we are *thrown* into, before anything like a decision is even possible. This is closely tied to the fact that interaction is primarily, that is, from the very beginning, embodied – a fact (or the facticity) of our physical nature, and specifically, of the kind of body that we have and the contingencies of our earliest existence. There is, in effect, no scientific mystery to this phenomenon, even if in everyday experience it seems a mystery in terms of why for the most part we cannot help but engage in it. The second aspect involves the idea that strong interaction is irreducible to the individuals involved.

I start with a question related to the first aspect, namely the question of the origin of interaction. Merleau-Ponty points to the bodily nature of interaction with his concept of intercorporeity (Merleau-Ponty 1962, p. 352). What I want to suggest is that intersubjective interaction ultimately derives from a more primary intercorporeal interaction.

We know the principle from neuroscience, *movement influences morphology* (Edelman 1992, Sheets-Johnstone 1998): brain development *results* from the system as a whole adapting to new levels of organization at more peripheral levels, rather than the neurological developments unfolding to 'allow' increasing proprioceptive capacities (Van der Meer and Van der Weel 1995). Consider the variety of developmental processes that follow this principle. For example, there is good evidence that both (1) a primitive proprioceptive registration of one's bodily movement, and (2) a differentiation between self and non-self develop prenatally (see Gallagher 1996). For example, proprioceptors in the muscles (muscle spindles) first appear at 9 weeks gestational age (Humphrey 1964); parts of the vestibular system develop as early as the fourth month of gestation (Jouen and Gapenne 1995); and cortical connections necessary for body-schematic proprioceptive processes are in place by 26 weeks gestational age. In addition, the differentiation between self and non-self in the later-term fetus is evidenced across a number of studies of fetal behavioral reaction to various stimuli. In response to auditory stimuli, as early as 24 weeks gestational age, fetal heart rate changes; and after 25 weeks, the fetus responds by blinking its eyes or moving its limbs. Cortical response to such stimuli has been demonstrated in

premature infants between 24-29 weeks gestational age (Fifer and Moon 1988). Differential responsiveness in the late-term fetus, signals a preference for some sounds (such as the mother's voice) rather than others (DeCasper and Spence 1986). Bright light directed on the lower abdomen of the mother in the third trimester can elicit fetal eye blinks (Emory and Toomey 1988), and fetal facial movements prompted by music or voice may be indicative of a similar differential awareness (Birnholz 1988). And we know that what Aristotle called the most basic sense, the tactile sense, develops early in the fetus, with cortical pathways intact by 20-24 weeks gestational age, with a differential registration between self-touching and being touched even earlier (Glass 2005).

Even before the development of full-fledged proprioceptive and tactile senses, however, the fetus is already moving. At twelve weeks gestational age, there is evidence of spontaneous and repetitious movements – e.g., movement of the hand to the mouth occurs multiple times an hour from this time (De Vries *et al.* 1982; Tajani and Ianniruberto 1990). At ten weeks gestational age fetuses display structured bodily movements which they develop through habituation (Krasnegor *et al.* 1998); for example, regular mouth opening and closing, swallowing, and movement in response to stimuli such as the mother's laugh or cough.

> The first movements to occur are sideward bendings of the head. [...] At 9-10 weeks [gestational] age complex and generalized movements occur. These are the so-called general movements [...] and the startles. Both include the whole body, but the general movements are slower and have a complex sequence of involved body parts, while the startle is a quick, phasic movement of all limbs and trunk and neck. (Prechtl 2001)

Two kinds of movement are involved here: early fetal movement, which is spontaneous and repetitive and starts out as a reflex that unfolds genetically (De Vries *et al.* 1982); and early fetal movement that appears regulated and practiced – i.e., non-reflex (Krasnegor *et al.* 1998) – and that starts out as a response to stimuli. Setting aside the question of which of these come first, we can say that at some point in early fetal motility responsive movement comes along.[4] The question is: To what is this movement a response? What is the

[4] I note here a recent study by Zoia *et al.* (2007) on intentional or directed movement in the fetus. Zoia *et al.* examined kinematic patterns of foetal movements showing that at 22 weeks hand to mouth and hand to head movement involved straighter and more accurately aimed trajectories with acceleration and deceleration phases consistent with target size and sensitivity. Thus, «by 22 weeks of gestation the movements seem to show the recognizable form of intentional actions, with kinematic

origination of this movement that helps to set the train of development in motion? The answer is that this kind of movement is a reaction to the mother's bodily movement – a kind of intercorporeal interaction.

> It is likely that these earliest regulated movements, which are prior to proprioceptive capacity, are a response within and to, the maternal body in *her* regulated and habituated, body schematic movement. [...] Add to physical movement the regular maternal heart beat, digestion, and breathing and we can see that the intrauterine world is not only a moving but quite rhythmic or regulated animate world. (Lymer 2010, p. 230)

This is not yet *intersubjective* interaction (the mother may not even know she's pregnant this early; and there is no claim that the fetus is an experiencing subject), but it is an *intercorporeal* interaction – a non-conscious motor coupling between mother and fetus driven toward and then driven by proprioception and touch. The point I want to make here is that whatever the moment of the awakening of consciousness – whether that is prenatal (at around 26 weeks gestational age) or later than that – and wherever we might locate the earliest aspect of self-awareness, this kind of intercorporeal interaction predates that, so that we find ourselves already immersed in interactive processes that prefigure the intersubjective ones found in primary intersubjectivity.

To this immersion I want to add that the primary and secondary intersubjective interactions that we find in infancy are more than capacities or mechanisms that belong to the individuals involved in interaction. They are not based simply on "first-order *mechanisms*" (Buckner *et al.* 2009) that we find in each individual, because they are not reducible to the sum of individual capacities (De Jaegher and Di Paolo 2007, De Jaegher *et al.* 2010). In the case of intersubjective interaction, $1 + 1 > 2$. This is what De Jaegher and Di Paolo (2007) mean by saying that interaction has some degree of autonomy. The interaction in intersubjective contexts goes beyond each participant; it results in something (the creation of meaning) that goes beyond what each individual qua individual can bring to the process – just as when two people dance the tango, something dynamic is created that neither one could create on their

patterns that depend on the goal of the action, suggesting a surprisingly advanced level of motor planning» (Zoia *et al.* 2007, p, 217). Also see Becher 2004 : «Purposive movement depends on brain maturation. This begins at about 18 weeks' gestation and progressively replaces reflex movements, which disappear by about 8 months after birth [...]»

own. Moreover, as we have just seen in regard to the origins of interaction, we are in the tango before we even know it.

So, not just in its origins, but as an ongoing process, interaction has a certain kind of irreducibility; it goes beyond the individual participants. In cases were one person is totally in control of the other person (if total control is ever possible), there is no interaction in this specific sense. The characteristics of immersion and irreducibility motivate the question about individual autonomy – self-agency. Merleau-Ponty talks about the infant getting caught up in the "whirlwind of language" – but prior to that the infant is caught up in the whirlwind of interaction – and even as adults we remain in that whirlwind. And within that whirlwind, does the irreducibility of interaction leave any room for self-agency or individual autonomy? If I, always already, even before birth, am caught up in a whirlwind of interaction, and that interaction always goes beyond me and my ultimate control, is there really any room for self-agency?

Self-Agency and the Narrative Self

There are current lively debates about self-agency and related concepts of freedom, free will, intention formation, and the sense of agency, with a variety of positions being staked out. From materialist and reductionist perspectives numerous theorists argue that self-agency is an illusion. They point to neuroscientific data (e.g., the Libet experiments that seem to show that the brain knows what we are going to do before we, as conscious individuals, do) or to the results of psychological experiments (e.g., Wegner 2002, Pockett 2006); or they suggest that if we do have free will, we need a subpersonal explanation of it that shows how it is generated in the individual brain (Spence 1996). Those who defend free will also often appeal to processes that are *in the head* (intention formation, reflective decision-making, or the phenomenological sense of agency, e.g., Pacherie 2008, Stephens and Graham 2000), or to mental causation, (Searle 1983, Lowe 1999). These approaches – whether they dismiss or defend the notion of free will – follow a traditional view that conceives of self-agency (or the lack of it) as a matter of individual subjectivity. Free will is either in the individual system or it is not. Even those theories that take social phenomena into account often use the individual as a measure of whether free will exists. For example, social

determinists argue that individual free will doesn't exist precisely because the individual is fully determined by our social interactions, cultural forces, etc.

In general, then, discussions of freedom/free will/self-agency focus on the individual – the question is framed that way, for example, if we ask about individual autonomy. I want to suggest, however, that in response to the question about self-agency, motivated by the account of strong interaction, we can conceive of self-agency in different terms by conceiving of the agent as something other than an individual who either has or does not have free will. If we view the agent as someone who emerges from intercorporeal interactions, and develops in social interactions with others, then we have a good model for speaking about self-agency in a system that is not reducible to a simple individual. On this model, self-agency – and a proper sense of freedom (which comes along with a proper sense of responsibility) – can be found only in the context of social interaction, where our intentions are formed in or out of our interactions with others.

Clearly, we learn to act from watching and interacting with others as they act in the world. We learn our own action-possibilities from others. Through our interactions with others we generate shared intentions and form our own intentions out of the same fabric. In this context, how can we explain self-agency?

It is at this point that I want to point out the importance of that aspect of interaction theory that involves communicative and narrative competencies (Gallagher and Hutto 2008). Beyond the processes of primary and secondary intersubjectivity, communicative and narrative practices allow for a certain volitional space to open up – the possibility of taking a critical perspective on ourselves. Narrative allows us to reflectively locate our interactions in what Bruner calls the 'landscape of action' and 'the landscape of consciousness' (Bruner 1986). That is, through narrative, we can reflect on our actions and interactions, and on what our motives for such actions might be.

In this process, and specifically in autobiographical (or self-) narrative, narrative distance, a concept that goes back to Aristotle's *Poetics*, is established between the self who narrates and the self who is narrated. This distance allows for the possibility of what Harry Frankfurt (1971) calls second-order volitions – that is, volitions in which we consider or evaluate our own first-order action volitions. On Frankfurt's view, this capacity for second-order volitions, or what Charles Taylor (1989) calls the possibility for a *strong evaluation* of our own desires, is essential for moral personhood. From an

interactionist perspective, this is possible only as a result of social interaction processes, in social settings where we act and interact, and where we exercise our communicative and narrative practices.

What we call autonomy, then, is not constituted in just an internal intra-individual negotiation made by an agent with respect to herself, but is inextricably interwoven into and out of our relationships with others. In this regard, self-agency becomes a matter of degree rather than an all or nothing issue. Some people arrange their lives with others, or find themselves in such arrangements, so that they have a high degree of freedom – a greater range of possibilities than others who find themselves in social relationships, or cultures, or institutions where they are prevented from acting freely.

There is nothing new in this thought: our social interactions and arrangements are such that they either promote freedom or prevent it. Whatever self-agency is, it's weaved out of this fabric of interaction; not a characteristic of the individual; but a characteristic of a set of relationships. In some of my interactions I am freer than in others. Some arrangements support self-agency, and some do not. I could say, without contradiction, that I am free and I am not free – but only in the sense that my self-agency is constituted in my different relations *differently*.

It's also the case that certain interactions can make one participant free and the other a slave. So the question that derives directly from conceiving of intersubjective interaction as a primary force in shaping our cognitive, emotional, and social life is not the *metaphysical* question: Do I as an individual have free will? It is rather the *political* question: who is free (or more free) and who is not, and why? The political question is a pragmatic and critical one, because we can ask why, and motivate change.

Conclusion

With regard to discussions of social cognition, shifting away from theory-of-mind approaches, such as theory theory and simulation theory, and taking up the interaction theory and the emphasis on intersubjective interaction also involves shifting away from conceptions of self-agency that are reducible to neural or mental or strictly individual processes framed in terms of mental causation. I've suggested that self-agency is a matter of degree and that it can be won or lost in the varying contexts of interaction – contexts from which I can

distance myself through a narrative process that allows for strong evaluation. Accordingly, self-agency and related phenomena such as free will and intention formation – these are not things that pertain strictly to an individual; rather, they are constituted in interaction and in communicative and narrative practices.

REFERENCES

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behavior: Action parsing and intentional inference. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, (pp. 193-206). Cambridge, MA: MIT Press.

Baldwin, D. A., & Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Science, 5*(4), 171-178.

Banks, W., Pockett, S., & Gallagher, S. (2006). *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition*, (pp. 109-124). Cambridge, MA: MIT Press.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, (pp. 233-251). Cambridge, MA: Basil Blackwell.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Becher, J.-C. (2004). Insights into early fetal development. *Behind the Medical Headlines*. (Royal College of Physicians of Edinburgh and Royal College of Physicians and Surgeons of Glasgow October 2004).

Birnholz, J. C. (1988). On observing the human fetus. In W. P. Smotherman & S. R. Robinson (Eds.), *Behavior of the Fetus*, (pp. 47-60). Caldwell, NJ: Telford Press.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Buckner, C., Shriver, A., Crowley, S., & Allen, C. (2009). How 'weak' mindreaders inherited the earth. *Behavioral and Brain Sciences, 32*(2), 140-141.

Catmur, C., Walsh, V., & Heyes, C. (2007). Sensorimotor learning configures the human mirror system. *Current Biology, 17*(17), 1527-1531.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148-153.

Currie, G. (2008). Some ways of understanding people. *Philosophical Explorations, 11*(3), 211-218.

De Casper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development, 9*, 133-150.

De Jaegher, H. (2008). Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition, 18*(2), 535-542.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences, 6*(4), 485-507.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Does social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(10): 441-447.

De Vries, J. I. P., Visser, G. H. A., & Prechtl, H. F. R. (1982). The emergence of fetal behaviour: I. Qualitative aspects. *Early Human Development, 7*, 301-322.

Dinstein, I., Thomas, C., Behrmann, M., & Heeger, D. J. (2008). A mirror up to nature. *Current Biology, 18*(1), R13-R18.

Dittrich, W. H., Troscianko, T., Lea, S. E. G., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25*, 727-738.

Edelman, G. (1992). *Bright Air, Brilliant Fire*. New York: Basic Books.

Emory, E. K., & Toomey, K. A. (1988). Environmental stimulation and human fetal responsivity in late pregnancy. In W. P. Smotherman & S.

R. Robinson (Eds.), *Behavior of the Fetus*, (pp. 141-161). Caldwell, NJ: Telford Press.

Fifer, W. P., & Moon, C. (1988). Auditory experience in the fetus. In W. P. Smotherman & S. R. Robinson (Eds.), *Behavior of the Fetus*, (pp. 175-188). Caldwell, NJ: Telford Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5-20.

Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences, 8*(4), 465-486.

Gallagher, S. (1996). The moral significance of primitive self-consciousness. *Ethics, 107*(1), 129-140

Gallagher, S. (2005). *How the Body Shapes the Mind.* Oxford: Oxford University Press/Clarendon Press

Gallagher, S. (2007). Simulation trouble. *Social Neuroscience, 2*(3-4), 353-365.

Gallagher, S. (in press). Narrative competency and the massive hermeneutical background. In P. Fairfield (Ed.), *Hermeneutics in Education*. New York: Continuum.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. Racine, C. Sinha & E. Itkonen (Eds.), *The Shared Mind: Perspectives on Intersubjectivity*, (pp. 17-38). Amsterdam: John Benjamins.

Gergely, G. (2001). The obscure object of desire: 'Nearly, but clearly not, like me': Contingency preference in normal children versus children with autism. *Bulletin of the Menninger Clinic, 65*(3), 411-426.

Glass, P. (2005). The vulnerable neonate and the neonatal intensive care environment. In G. B. Avery, M. G. MacDonald, M. M. K. Seshia & M. D. Mullett (Eds.), *Avery's Neonatology: Pathophysiology & Management of the Newborn*, (pp. 111-129). Philadelphia: Lippencott, Williams and Wilkins.

Goldman, A. I. (1995). Desire, intention and the simulation theory. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, (pp. 207-224). Cambridge, MA: MIT Press.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories.* Cambridge, MA: MIT Press.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language, 1*(2), 158-171.

Gordon, R. M. (1995). Developing commonsense psychology: Experimental data and philosophical data. Paper presented at the APA Eastern Division Symposium on Children's Theory of Mind, 12/27/95. (http://www.umsl.edu/~philo/Mind_Seminar/New%20Pages/papers/Gordon/apakids9.htm)

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, Mind, and Logic*, (pp. 45-59). Cambridge: Cambridge University Press.

Heal, J. (1998a). Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language, 13*, 477-498.

Heal, J. (1998b). Understanding other minds from the inside. In A. O'Hear (Ed.), *Current Issues in Philosophy of Mind*. Cambridge: Cambridge University Press.

Hobson, P. (1993). The emotional origins of social understanding. *Philosophical Psychology, 6*(3), 227-249.

Hobson, P. (2002). *The Cradle of Thought*. London: Macmillan.

Hobson, P., & Lee, A. (1999). Imitation and identification in autism. *Journal of Child Psychology and Psychiatry, 40*(4), 649-659.

Humphrey, T. (1964). Some correlations between the appearance of human fetal reflexes and the development of the nervous system. *Progress in Brain Research, 4*, 93-135.

Issartel, J., Marin, L., & Cadopi, M. (2007). Unintended interpersonal coordination: 'Can we march to the beat of our own drum'? *Neuroscience Letters, 411*(3), 174-179.

Jouen, F., & Gapenne, O. (1995). Interactions between the vestibular and visual systems in the neonate. In P. Rochat (Ed.), *The Self in Infancy: Theory and Research*, (pp. 277-301). Amsterdam: Elsevier Science.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge: Cambridge University Press.

Keysers, C., & Gazzola, V.(2006). Towards a unifying neural theory of social cognition. In S. Anders, G. Ende, M. Junghofer & J. Kissler (Eds.), *Understanding Emotions*, (pp. 379-402). Amsterdam: Elsevier.

Krasnegor, N. A., Fifer, W., Maulik, D., McNellis, D., Romero, R., & Smotherman, W. (1998). Fetal behavioral development: Measurement of habituation, state transitions, and movement to assess fetal well being and to predict outcome. *The Journal Maternal-Foetal Investigation, 8*, 51–57.

Leslie, A. (1991). The theory of mind impairment in autism: Evidence for a modular mechanism of development? In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, (pp. 63-78). Oxford: Blackwell.

Lindblom, J. (2007). Minding the body: Interacting socially through embodied action. *Linköping Studies in Science and Technology*. Dissertation No. 1112.

Lindblom, J., & Ziemke, T. (2007). Embodiment and social interaction: Implications for cognitive science. In T. Ziemke, J. Zlatev & R. Frank (Eds.), *Body, Language, and Mind: Embodiment*, (pp.129-162). Berlin: Mouton de Gruyter.

Lowe, E. J. (1999). Self, agency and mental causation. *Journal of Consciousness Studies, 6*(8-9), 225-239.

Lymer, J. (2010). The phenomenology of the maternal-foetal bond. Ph.D. Dissertation. Wollongong University, Australia.

Maurer, D., & Barrera, M. E. (1981). Infants' perception of natural and distorted arrangements of a schematic face. *Child Development, 52*(1), 196-202.

Meltzoff, A., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75-78.

Meltzoff, A., & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development, 17,* 83-99.

Merleau-Ponty, M. (1962). *Phenomenology of Perception.* (Tr. by C. Smith). London: Routledge and Kegan Paul.

Muir, D., & Lee, K. (2003). The still-face effect: Methodological issues and new applications. *Infancy, 4*(4), 483-491.

Murray, L., & Trevarthen, C. (1985). Emotional regulation of interactions between 2-month-olds and their mothers. In T.M. Field & N. A. Fox (Eds.), *Social Perception in Infants*, (pp. 177-197). Norwood, NJ: Ablex.

Nadel, J., Croué, S., Mattlinger, M.-J, Canet, P., Hudelot, C., Lécuyer, C, & Martini, M. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science, 2*, 164-173.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford: Clarendon Press.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*(1), 179–217.

Prechtl, H. (2001). Prenatal and early postnatal development of human motor behavior. In A. F. Kalverboer & A. Gramsbergen (Eds.), *Handbook of brain and behaviour in human development*, (pp. 415-418). Dordrecht: Kluwer Academic Publishers.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review Neuroscience, 2*, 661-670.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Senju, A., Johnson, M. H., & Csibra, G. (2006). The development and neural basis of referential gaze perception. *Social Neuroscience, 1*(3-4), 220-234.

Sheets-Johnston, M. (1998). Consciousness: A Natural History. *Journal of Consciousness Studies, 5*(3), 260-294.

Spence, S. A. (1996). Free will in the light of neuropsychiatry. *Philosophy, Psychiatry, and Psychology, 3*(2), 75-90.

Stephens, G. L., & Graham, G. (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, MA: MIT Press.

Stormark, K. M., & Braarud, H. C. (2004). Infants' sensitivity to social contingency: A "double video" study of face-to-face communication between 2- and 4-month-olds and their mothers. *Infant Behavior and Development, 27*, 195-203.

Tajani, E., & Ianniruberto, A. (1990). The uncovering of fetal competence. In M. Papini, A. Pasquinelli & E. A. Gidoni (Eds.), *Development Handicap and Rehabilitation: Practice and Theory*. Amsterdam: Elsevier Science Publishers.

Taylor, C. (1989). *Sources of the Self*. Cambridge, MA: Harvard University Press.

Trevarthen, C. B. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before Speech*, (pp. 321-348). Cambridge: Cambridge University Press.

Trevarthen, C., & Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock (Ed.), *Action, Gesture and Symbol: The Emergence of Language*, (pp. 183-229). London: Academic Press.

Tronick, E. Z. (2007). *The Neurobehavioral and Social Emotional Development of Infants and Children*. New York: Norton.

Tronick, E., Als, H., Adamson, L.,Wise, S., & Brazelton, T. B. (1978). The infants' response to entrapment between contradictory messages in

face-to-face interactions. *Journal of the American Academy of Child Psychiatry, 17*, 1-13.

Van der Meer, A. L., Van der Weel, F. R., & Lee, D. N. (1995). The functional significance of arm movements in neonates. *Science, 267*, 693-695.

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology, 33*, 514-535.

Wegner, D. (2002). *The Illusion of Conscious Will.* Cambridge, MA: MIT Press.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*, 73–77.

Zoia, S., Blason, L., D'Ottavio, G., Bulgheroni, M., Pezzetta, E., Scabar, A., Castiello, U. (2007). Evidence of early development of action planning in the human foetus: A kinematic study. *Experimental Brain Research, 176*(2), 217-226.

The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits*

Terry Horgan **
thorgan@email.arizona.edu

ABSTRACT

The paper will focus on three interrelated matters. First is the phenomenology of agency, the "what it is like" of experiencing oneself as an agent – and more specifically, the experiential aspect of freedom that is an integral part of the phenomenology of agency. Second is the extent to which introspection is, or is not, a reliable way to answer questions about the phenomenology of agency and freedom. Third is the import of these first two matters for philosophical debates about agency and free will.

Briefly, my overall position goes as follows. The phenomenology of free agency has features that are well and aptly described by language of the kind that is traditionally employed by advocates of metaphysical libertarianism concerning the free will issue – language like "self as ultimate source," and "agent as cause". This is something that is reliably detectable by introspection. However, introspection by itself cannot reliably ascertain whether or not the satisfaction conditions for free-agency phenomenology require, for example, the falsity of state-causal determinism or the presence of the metaphysically heavyweight attribute that metaphysical libertarians call "agent-causal freedom".

Moreover, the best overall theoretical position about the nature of free agency – the one that emerges by abductive "inference to the best explanation" of all pertinent, evidentially relevant, factors – is compatibilist. Among the considerations that underwrite this abductive conclusion is the fact that a suitable version of compatibilism can provide a full accommodation of the phenomenology of free agency; i.e., the right kind of compatibilism entails

* My thanks to Michael McKenna and Mark Timmons for helpful comments and discussion.
** University of Arizona

that normal humans do indeed exercise free agency, and also entails that their agentive experience does not misrepresent the nature of free agency itself.

---

1. Some Reliably Introspectible Aspects of Agentive Phenomenology[2]

I begin by describing some features of agentive phenomenology which, I submit, are readily ascertainable just on the basis of introspective attention to such phenomenology.

What is behaving like phenomenologically, in cases where you experience your own behavior as action? Suppose that you deliberately do something – say, holding up your right hand and closing your fingers into a fist. What can you ascertain about the phenomenology of this item of behavior, on the basis of introspective attention to this phenomenology? To begin with, there are of course the purely bodily-motion aspects of the phenomenology – the what-it's-like of being visually and kinesthetically presented with your own right hand rising and its fingers moving into clenched position. But there is more to it than that, of course, because you are experiencing this bodily motion *as your own action*.

In order to help bring into focus this specifically actional phenomenological dimension of the experience, it will be helpful to approach it a negative/contrastive way, via some observations about what the experience is *not* like. For example, it is certainly not like this: first experiencing an occurrent wish for your right hand to rise and your fingers to move into clenched position, and then passively experiencing your hand and fingers moving in just that way. Such phenomenal character might be called *the phenomenology of fortuitously appropriate bodily motion*. It would be very strange indeed, and very alien.

Nor is the actional phenomenological character of the experience like this: first experiencing an occurrent wish for your right hand to rise and your fingers to move into clenched position, and then passively experiencing a causal process consisting of this wish's causing your hand to rise and your fingers to move into clenched position. Such phenomenal character might be called *the*

[2] This section is adapted, with some modifications, deletions, and additions, from similar sections in Horgan 2007b and in Horgan *et al.* 2003.

passive phenomenology of psychological state-causation of bodily motion.[3] People often do passively experience causal processes *as* causal processes, of course: the experience of seeing the collision of a moving billiard ball with a motionless billiard ball is an experience as-of the collision causing the latter ball's subsequent motion; the experience of observing the impact of the leading edge of an avalanche with a tree in its path is an experience as-of the impact causing the tree to become uprooted; and so on. Sometimes people even experience their own bodily motions as state-caused by their own mental states – e.g., when one feels oneself shuddering and experiences this shuddering as caused by of a state of fear. But it seems patently clear that one does not normally experience one's own actions in that way – as passively noticed, or passively introspected, causal processes consisting in the causal generation of bodily motion by occurrent mental states. That too would be a strange and alienating sort of experience.[4]

How, then, should one characterize the actional phenomenal dimension of the act of raising one's hand and clenching one's fingers, given that it is not the phenomenology of fortuitously appropriate bodily motion and it also is not the passive phenomenology of psychological state-causation of bodily motion? Well, it is the what-it's-like of *self as source* of the motion. You experience your arm, hand, and fingers as being moved *by you yourself*–rather than experiencing their motion either as fortuitously moving just as you want them to move, or passively experiencing them as being caused by your own mental states. You experience the bodily motion as generated by *yourself.*

The language of causation seems apt here too, but differently deployed: you experience your behavior as *caused* by you yourself, rather than experiencing it as caused by *states* of yourself. Metaphysical libertarians about human freedom sometimes speak of "agent causation" (or "immanent causation"), and such terminology seems *phenomenologically* apt regardless of what one thinks about the intelligibility and credibility of metaphysical libertarianism. Chisholm (1964) famously argued that immanent causation (as he called it) is a distinct species of causation from event causation (or "transeunt" causation, as he called it). But he later changed his mind (Chisholm 1995), arguing instead

[3] Here and throughout I speak of 'state-causation' rather than 'event-causation'. More below on my reasons for this choice of terminology. States can be short-lived, and often when they do they also fall naturally under the rubric 'event.'

[4] For discussion of a range of psychopathological disorders involving similar sorts of dissociative experience, see Stephens and Graham (2000).

that agent-causal "undertakings" (as he called them) are actually a species of event-causation themselves – albeit a very different species from ordinary, nomically governed, event causation. Phenomenologically speaking, there is indeed something episodic – something temporally located, and thus "event-ish" – about experiences of self-as-source. Thus, the expression 'state causation' works better than 'event causation' as a way of expressing the way behaviors are *not* presented to oneself in agentive experience. Although agentive experience is indeed "event-ish" in the sense that one experiences oneself as undertaking to perform actions *at specific moments in time*, one's behavior is not experienced as caused by *states* of oneself.

The phenomenology of doing typically includes another aspect which will be especially important in the context of the present paper: what I will call *core optionality*. (More presently on the reason for the modifier 'core'). Normally when you do something, you experience yourself as *freely* performing the action, in the sense that it is *up to you* whether or not to perform it. You experience yourself not only as generating the action, and not only as generating it purposively, but also as generating it in such a manner that you *could have done otherwise*. This palpable phenomenology of optionality has not gone unrecognized in the philosophical literature on freedom and determinism, although often in that literature it does not receive as much attention as it deserves. (Sometimes the most explicit attention is given to effort of will, although it takes only a moment's introspection to realize that the phenomenology of effortfully exerting one's will is really only one, quite special, case of the much more pervasive phenomenology of optionality[5]).

The core-optionality aspect of agentive phenomenology is intimately bound up with the aspect of self-as-source, in such a way that the former is an essential component of normal agentive self-source experience.[6] In experiencing one's

[5] This is not to deny, of course, that there is indeed a distinctive phenomenology of effort of will that *sometimes* is present in the phenomenology of doing. The point is just that this aspect is not always present. A related phenomenological feature, often but not always present, is the phenomenology of *trying* – which itself is virtually always a dimension of the phenomenology of effort of will, and which often (but not always) includes a phenomenologically discernible element of uncertainty about success. (Sometimes the phenomenological aspect of core optionality attaches mainly to the trying dimension of the phenomenology of doing. When you happen to succeed at what you were trying to do but were not at all confident you could accomplish – e.g., sinking the 10 ball into the corner pocket of the pool table – the success aspect is not experienced as something directly under voluntary control).

[6] I say that the aspect of core optionality is an essential component of *normal* self-source experience because I mean to leave open the possibility of unusual self-source experiences that lack

behavior as emanating from oneself as its source, one experiences oneself as being able to refrain from so behaving – or at any rate, as being able to refrain from willfully producing such behavior. This is so even when acts under extreme coercion or duress – e.g., handing over one's wallet or purse to a thief who is pointing a gun in one's face. It also is so even when one acts with an extreme phenomenological "imperativeness"– e.g., a mother's unhesistatingly leaping into the river to save her drowning child, Luther's acting out a sense of moral requirement (as expressed by his declaring "Here I stand, I can do no other"), the compulsive hand-washer's act of washing hands for the third time in ten minutes. The *core* phenomenology of optionality that is essential to ordinary agentive experience remains present in all such cases, even though there are further, superimposed, phenomenological aspects (duress, moral-obligation experience, intensely strong irrational desires, or the like) whose presence can render appropriate, in context, a judgment that the agent "could not have done otherwise," or "had no other option," or "did not act freely". Because the phenomenology of core optionality remains present even in such cases, it also can be contextually appropriate to use 'could' and 'option' and 'free' in a way that reflects this fact (rather than in a way that reflects the presence of one or another kind of superimposed non-optionality phenomenology). For instance, one might say this:"I could have refrained from giving the thief my wallet, and thus I gave it to him freely and with the option of refraining – even though refraining would have been quite stupidly irrational". Hereafter I will use the expression 'free-agency phenomenology', in order to refer to the experience of self-as-source in a way a way that underscores the aspect of core optionality that is an essential component of normal self-as-source experience.

A few words are in order at this point about thought-experimental "Frankfurt scenarios" inspired by Frankfurt (1969). One such scenario is this: one's body would have moved the same way even if one had not willed it to move that way, because a device implanted in one's motor cortex would have triggered that same motion had one not willfully produced it; but in that case the motion would not have been experienced as willfully generated, and indeed

this aspect – for instance, self-source experiences in which one firmly believes that one is in a "Frankfurt scenario" in which one's circumstances are such that were one about to will to refrain from performing the act one is about to perform, an evil scientist would cause the pertinent bodily motions to occur anyway and would also cause these motions to be accompanied by (epiphenomenal) experience-as-of willfully performing that action. More momentarily on Frankfurt scenarios.

would not have been experienced as one's own action. A different Frankfurt scenario is this: one's body would have moved the same way even if one had not willed it to move that way, because a device implanted in one's motor cortex would have triggered that same motion had one not willfully produced it; in addition, that device would have triggered the phenomenology of willing to move one's body in just that way – with the dual triggering operating in a manner that renders the phenomenology itself completely epiphenomenal vis-à-vis the bodily motion. As far as free-agency *phenomenology* is concerned (and that is the present topic), the main thing to stress is the following: in both of these scenarios (and in most Frankfurt-style scenarios), one's free-agency phenomenology is at least partially *non-veridical*, because the phenomenology includes not only the self-as-source aspect but also the could-do-otherwise aspect that is an essential component of normal self-as-source experience. The agent's phenomenology is as-of being a *full-fledged* self-source of the behavior, where full-fledgedness includes being such that one could have acted otherwise instead; but in Frankfurt scenarios, the agent is not a full-fledged self-source of the kind that the agent experiences himself/herself to be.[7] These remarks about agentive phenomenology leave various moral and metaphysical questions still open – e.g., (i) whether the agent in a Frankfurt scenario is morally responsible for the action, (ii) whether the agent is a genuine self-source of the behavior even though the agent could not have done otherwise, and (iii) whether the agent acts freely even though the agent could not have done otherwise. Whatever one might say about those questions, the key point is that the self-as-source aspect of normal agentive *experience* includes the core optionality (core "could-do-otherwise") aspect as an essential element.

Agentive phenomenology is more closely akin to perceptual/kinesthetic experience than it is to discursive thought. (Many higher non-human animals, I take it, have some agentive phenomenology, even if they engage in little or no discursive thought). Of course, we humans also wield *concepts* like agency, voluntariness, and the like (whereas it is questionable whether non-human

[7] What about the Frankfurt scenario envisioned in note 6, in which one firmly believes that one is in a scenario in which core optionality is absent? Perhaps here one's agentive phenomenology would be as-of *non*-full-fledged self-as-source-hood in which the core-optionality aspect is lacking. But that would be extremely unlike ordinary agentive phenomenology. (Alternatively – as I myself suspect would be the case – perhaps even here the core-optionality aspect still would be present in one's agentive *phenomenology* despite one's *belief* that core optionality itself is absent. Compare experiences of the Muller-Lyer illusion, in which one horizontal line still looks longer than the other even when one firmly believes the two lines are the same length).

animals do); but thoughts employing these concepts are not to be conflated with agentive phenomenology itself.

2. Some Limitations of Introspection Vis-à-vis Agentive Phenomenology[8]

The phenomenal character of one's current experience is self-presenting to the experiencing subject. Self-presentingness is an especially intimate form of direct acquaintance between the experiencing subject on one hand, and the phenomenal character of some aspect of the subject's current state of phenomenal consciousness; the state's appearing a certain way, acquaintance-wise, is constitutive of the state's actually being that way.

Let a *purely phenomenological question* be a question that (i) is about some aspect of the intrinsic phenomenal character of one's present experience, and (ii) is such that the answer is entirely determined just by the intrinsic phenomenal character of one's present experience. (The point of clause (ii) is to exclude questions that bring in some extrinsic aspect while still being in some sense "about" intrinsic phenomenal character – e.g., "Am I now undergoing an experience with the phenomenal character that I was writing about last Tuesday?").

In light of the fact that phenomenal character is self-presenting, one might be tempted to think that any purely phenomenal question can be reliably answered directly on the basis of introspection. More specifically, one might be tempted to think that introspection alone can reliably determine whether or not free-agency phenomenology has metaphysical-libertarian satisfaction conditions. (Having metaphysical-libertarian satisfaction conditions means this: the intentional content of one's free-agency experience is veridical only if one is an "agent-cause" in the metaphysically heavyweight sense of this notion that is invoked by metaphysical libertarians – which entails, inter alia, that state-causal determinism is false).

I maintain, however, that this claim about the powers of introspection vis-à-vis free-agency phenomenology is false. (Hence the more general thesis – that

[8] This section is adapted, with some modifications and deletions, from section 3 of Horgan (in press-b). Other pertinent discussions of mine, sometimes collaborative, are Horgan (2007a, in press-a) and Horgan and Timmons (in press).

any purely phenomenal question can be reliably answered directly on the basis of introspection – is also false). In this section I will briefly say why.[9]

Let me begin by introducing some terminology. First, I distinguish two kinds of introspection concerning one's current experience. On one hand is *attentive* introspection: paying attention to certain aspects of one's current experience. On the other hand is *judgmental* introspection: the process of forming a judgment about the nature of one's current experience, and doing so spontaneously just on the basis of attending to the aspects(s) of one's current experience about which one is judging – without any reliance on collateral information or evidence. (Judgmental introspection thus deploys attentive introspection, while also generating a judgment about what is being attended to).

Second, I call a purely phenomenal question *conceptual-competence amenable* (for short, CC amenable) just in case it can be correctly answered by simply introspectively attending to one's current experience and then spontaneously exercising one's conceptual competence with the pertinent concepts. By contrast, a purely phenomenal question is *conceptual-competence transcendent* (for short, CC transcendent) just in case it *cannot* be correctly answered this way.

With these distinctions at hand, consider now the following three pairwise-incompatible claims about the satisfaction conditions of free-agency phenomenology.

(1) Free-agency phenomenology has satisfaction conditions that (i) are fully fixed by intrinsic phenomenal character alone, and (ii) are metaphysical libertarian.

(2) Free-agency phenomenology has satisfaction conditions that (i) are fully fixed by intrinsic phenomenal character alone, and (ii) are compatible with state-causal determinism (and hence are not metaphysical-libertarian).

(3) Free-agency phenomenology has satisfaction conditions that (i) are not fully fixed by phenomenal character alone, (ii) instead are fixed by phenomenology in combination with extra-phenomenological facts about the experiencing agent's cognitive architecture, and (iii) are

[9] For more extended elaboration and defense of the view, see Horgan (in press-b).

such that their being metaphysical-libertarian or not, and their being compatible with state-causal determinism or not, depends upon those cognitive-architecture facts.

Claims (1) and (2) both construe free-agency phenomenology as having "purely narrow" referential purport that lacks any constitutive externalistic elements, whereas claim (3) construes it as having "wide" referential purport that incorporates certain constitutive externalistic elements. For the phenomenology to have wide referential purport is for its reference-relation to its referent-property (if it has a referent-property) to depend constitutively not merely on the intrinsic character of the phenomenology itself, but also upon certain phenomenology-external facts about the nature of the experiencing agent – according to claim (3), facts about the agent's cognitive architecture. On one potential view that comports with claim (3), the pertinent facts would concern the nature of the cognitive-architectural choice-generating and behavior-generating mechanisms that are normally operative in situations where the experiencing agent undergoes free-agency phenomenology, and meeting the satisfaction conditions would be a matter of exercising those cognitive mechanisms in the normal way.

Claims (1) and (2), on the other hand, construe free-agency phenomenology as referring, in the experience of all actual and possible creatures who are phenomenal duplicates of one another, to one and the same property – regardless of any differences in the cognitive architectures of different phenomenal duplicates.[10] The essence of the property that constitutes free agency is entirely fixed by the intrinsic phenomenal character of free-agency experience alone. Claim (1) says that this phenomenologically fixed property has metaphysical-libertarian satisfaction conditions, whereas claim (2) says that it has satisfaction conditions that are compatible with state-causal determinism (and hence are not metaphysical-libertarian).

Consider now the following question, which pertains entirely to the intrinsic phenomenal character of agentive experience and whose answer depends only on that phenomenal character – and which is therefore a purely phenomenological question:

[10] This property need not actually be instantiated by the creature in order to be the referent-property of the creature's free-agency experience. Indeed, it need not even be a property whose instantiation is metaphysically possible. (Maybe it is a metaphysical-libertarian property, and maybe – as some hard incompatibilists maintain – the instantiation of such a property is outright impossible regardless of whether or not state-causal determinism is true).

(Q) Which (if any) of the pairwise incompatible claims (1)-(3) is correct?

At the moment, the issue I am focusing upon is not what the answer is to question (Q), but rather this: whether or not one can reliably ascertain, just via judgmental introspection, what the answer is. I claim that one *cannot* do so, and that the reason why not is that (Q) is a CC transcendent question. Elsewhere (Horgan, in press-b) I defend these claims, and I also offer a proposed multi-component debunking explanation of the common judgmental-introspective beliefs that (a) one *can* reliably answer question (Q) just on the basis of introspection, and (b) that the answer is that claim (1) is true.

An explanatory task arises at this point that needs addressing – viz., the task of explaining credibly *why it should be* that (Q) is a CC-transcendent question. Since claims (1)-(3) all concern only the phenomenal character of free-agency experience, and since phenomenal character is self-presenting to the experiencing agent, something needs saying about why human agents are nonetheless unable to "read off" the answer to question (Q) just by directing their attentive introspection upon their own free-agency experience and then exercising their conceptual competence with concepts like the concept of state-causal determinism and the concept of free-agency phenomenal character.

I have addressed this explanatory task most extensively in in Horgan (in press-b); there is also pertinent discussion in Horgan (2007a, 2007b) and in Horgan and Timmons (in press). Although I lack the space here to rehearse my proposed account, let me just mention 3 key elements of the account. First, normal conceptual competence is mainly a matter of being able to correctly apply a given concept *to a concrete case* – or more precisely, to do so modulo one's available evidence; consequently, conceptual competence alone is apt to be fairly limited as a basis for answering abstract general questions about the nature of satisfaction conditions. Second, these same facts about conceptual competence are in play when one introspectively attends to one's agentive phenomenology with the goal in mind of forming an introspective judgment about question (Q): it is unreasonable and unwarranted to expect one's capacity for concept-wielding to be that splendid when it is directed at general hypotheses concerning the intentional content of agentive phenomenology, just as it is unreasonable to expect it to be that splendid when it is directed at general hypotheses concerning the satisfaction conditions for concepts themselves. Hence third, general hypotheses about satisfaction conditions are

a matter for abductive inference – even when these hypotheses concern facts about the intentional content of self-presenting phenomenal character, facts that are fully fixed by that phenomenal character itself.

3. Lessons

Let me draw out some lessons of the above discussion, with respect to philosophical debates about free agency. To begin with, participants in these debates need to explicitly acknowledge the existence of free-agency phenomenology – including its self-as-source dimension, and including the core optionality (core can/could do otherwise) aspect that is itself an essential component of normal self-as-source experience.[11]

Second, it needs to be appreciated that there are intimate interconnections among these three matters: (1) the satisfaction conditions of free-agency phenomenology, (2) the satisfaction conditions of everyday statements and judgments that ascribe free agency or classify specific acts and decisions as the products of free agency, and (3) the metaphysics of free agency. Item (1) is apt to constrain item (2), in the following way: if free-agency phenomenology has metaphysical-libertarian satisfaction conditions, then thereby so do everyday ascriptions of free agency, whereas if free-agency phenomenology has compatibilist satisfaction conditions, then thereby so do everyday ascriptions of free agency. In addition, item (1) is apt to constrain item (3), as follows: if genuine free agency exists at all, then it fully conforms to the satisfaction conditions imposed on it by agentive phenomenology. (I will express these modes of constraint by saying that free-agency phenomenology *strongly constrains*, respectively, the concept of free agency and the metaphysics of free agency. And I will say that an overall position that treats the concept of free

[11] Some philosophers, notably Eddy Nahmias and has collaborators, do pay attention to free-agency phenomenology and yet deny that it really has an aspect of self-as-source. (See, e.g., Nahmias *et al.* 2004). But they appear to assume that if there were such an aspect, then (a) this aspect would have metaphysical-libertarian satisfaction conditions, and (b) its having metaphysical-libertarian satisfaction conditions would be reliably ascertainable introspectively. They thereby conflate two claims: (1) the claim that agentive phenomenology has a self-as-source aspect, and (2) the claim that agentive phenomenology has a self-as-source aspect with features (a) and (b). In my view they would be right to deny claim (2), but they are wrong to deny claim (1) – and they unfortunately muddy up the dialectical waters by conflating the two claims.

agency and the metaphysics of free agency as strongly constrained by free-agency phenomenology is a *strongly internally coherent* position).

Third, it is important to articulate various package-deal positions that simultaneously address items (1), (2), and (3), and it is important to subject such positions to comparative cost-benefit assessments *as* package deals. Concerning item (1), a package-deal position will embrace just one of these two (incompatible) claims: (1a) phenomenological libertarianism, asserting that free-agency phenomenology has metaphysical-libertarian satisfaction conditions, or (1b) phenomenological compatibilism, asserting that such phenomenology has compatibilist satisfaction conditions. Likewise, concerning item (2) there are two options: (2a) conceptual libertarianism, asserting that everyday free-agency ascriptions have metaphysical-libertarian satisfaction conditions, or (2b) conceptual compatibilism, asserting that such ascriptions have compatibilist satisfaction conditions. Concerning item (3) there are three options: (3a) metaphysical libertarianism, (3b) metaphysical compatibilism, or (3c) hard incompatibilism.

Fourth, barring powerful countervailing theoretical considerations, theoretical package-deal positions that are strongly internally coherent will be much more likely to be correct than those that are not. (The default theoretical presumptions are that free agency has the features it is experienced as having, and that the concept of free agency has satisfaction conditions that conform well to the satisfaction conditions of free-agency experience. People implicitly adopt these presumptions routinely, and people routinely implicitly take the presumptions to be epistemically well warranted. In principle, one could challenge these default presuppositions, but doing so in a credible way would require some heavy-duty, hard-to-envision, form of argumentation). A strongly internally coherent package-deal position have will have these two features: first, it embraces (1a) if and only if it embraces (2a), and it embraces (1b) if and only if it embraces (2b); second, it asserts that if there is such a genuine phenomenon as free agency at all, then that phenomenon conforms to the satisfaction conditions laid down by free-agency phenomenology.

The fifth moral is conditional: if one can reliably ascertain, just on the basis of introspection, that free-agency phenomenology has metaphysical-libertarian satisfaction conditions, then there are only two package-deal positions that are strongly internally coherent, viz., (1a) + (2a) + (3a), and (1a) + (2a) + (3c). The first of these embraces phenomenological libertarianism, plus conceptual libertarianism, plus metaphysical libertarianism. This package deal is

libertarian through and through. The second view embraces phenomenological libertarianism, plus conceptual libertarianism, plus hard incompatibilism. This package deal asserts that there is no such phenomenon as free agency, on the grounds that (i) genuine free agency would have to conform to metaphysical-libertarian satisfaction conditions, and (ii) no real phenomenon conforms to such conditions.

The sixth moral is also conditional, and is a corollary of the fifth one: if one can reliably ascertain, just on the basis of introspection, that free-agency phenomenology has metaphysical-libertarian satisfaction conditions, then there is no viable compatibilist package-deal position that is strongly internally coherent. Thus the best one could do, by way of formulating a package-deal position that honors the introspectively manifest fact that free-agency phenomenology has metaphysical-libertarian satisfaction conditions, would be to adopt a partial-error theory asserting that although there really is a phenomenon of free agency, the nature of this phenomenon is very significantly misrepresented by free-agency experience. That kind of view is a very unattractive theoretical option for those who are inclined to reject metaphysical libertarianism. One reason to think so, inter alia, is that whatever phenomenon the account ends up treating as the one picked out by free-agency experience will be so different in reality from how it is experienced to be that there will be very little credible basis for claiming that it is an eligible referent of free-agency phenomenology (or of the concept of free agency).[12]

The six morals lately mentioned all draw upon the discussion in section 1 above, concerning reliably introspectible aspects of free-agency phenomenology. Let us now factor in the discussion in section 2, concerning the limitations of introspection with respect to free-agency phenomenology. That discussion yields this seventh moral: it is not the case that one can reliably ascertain, just on the basis of introspection, what the answer is to question (Q). This in turn brings an eighth moral in its wake, as a corollary: viz., it is not the

[12] For taxonomic completeness, the following additional moral is worth mentioning, also conditional in form: if one can reliably ascertain, just on the basis of introspection, that free-agency phenomenology has *compatibilist* satisfaction conditions, then the only strongly internally coherent package-deal position that conforms with the introspectively ascertainable nature of free-agency phenomenology is package-deal compatibilism, i.e., (1b) + (2b) + (3b). But it is extremely implausible to claim that it is introspectively *obvious* that self-as-source phenomenology has compatibilist satisfaction conditions, and I know of no compatibilist who does claim this. Rather, compatibilists tend either to ignore free-agency phenomenology altogether (the more typical tendency), or else to deny that agentive phenomenology has a self-as-source aspect at all (as do Nahmias and his collaborators).

case that one can reliably ascertain, just on the basis of introspection, that free-agency phenomenology has metaphysical-libertarian satisfaction conditions.

This leads to a ninth moral: there is another package-deal position that is consistent with what is reliably introspectively ascertainable about agentive phenomenology – viz., the position (1b) + 2(b) + 3b). This view is thoroughly compatibilist – phenomenologically, conceptually, and metaphysically – and is therefore strongly internally coherent. It begins with the contention that free-agency phenomenology has compatibilist satisfaction conditions. It then claims that free-agency phenomenology constrains both the concept of free agency and the metaphysics of free agency – in such a way that the concept has compatibilist satisfaction conditions too, and in such a way that genuine free agency is a phenomenon that is compatible with state-causal determinism (and hence is not correctly characterized by metaphysical libertarianism).

A tenth moral, also grounded in my discussion in section 2 of the limitations of introspection, is that there is an important role for abduction when one inquires about the satisfaction conditions of free-agency phenomenology – a role that is complementary to the roles of attentive and judgmental introspection, and that potentially can take up the slack left by introspection. That is good news for compatibilists, myself included.

4. Sketch of a Version of Package-Deal Compatibilism

Let me now briefly sketch the version of package-deal compatibilism that I favor.[13] I have defended various aspects of this overall approach in a number of prior writings, some collaborative (Horgan 1979, 2007a, 2007b, in press-a, in press-b, Graham and Horgan 1994, Henderson and Horgan 2000, Horgan and Timmons in press). The argumentation in those writings is largely abductive, and incorporates the contention that one cannot reliably ascertain the satisfaction conditions of free-agency phenomenology just on the basis of careful introspection.[14]

[13] This section is adapted, with some modifications and deletions, from section 4 of Horgan (2007b).

[14] I believe that there is significant work yet to be done by way of further elaborating my recommended approach – in particular, there is a need to say more about the satisfaction conditions of free-agency phenomenology, and about why and how these conditions can be met even if state-causal determinism is true. I am unhappy with possible-worlds satisfaction conditions according to which the possible worlds that are "accessible" to a freely choosing/acting agent include worlds in which a

As a prelude, let me distinguish two kinds of mental intentionality, which I call *presentational* content and *judgmental* content, respectively. Presentational intentional content is the kind that accrues to phenomenology directly – apart from whether or not one has the capacity to articulate this content linguistically and understand what one is thus articulating, and apart from whether or not one has the kind of sophisticated conceptual repertoire that would be required to understand such an articulation. Judgmental intentional content, by contrast, is the kind of content possessed by such linguistic articulations, and by the judgments they articulate. (Here I use 'judgment' broadly enough to encompass various non-endorsing propositional attitudes, such as *wondering whether*, *entertaining that*, and the like). Dogs, cheetahs, and numerous other non-human animals presumably have agentive phenomenology with presentational intentional content, although it is plausible that they have little or no sophisticated conceptual capacities of the kind required to undergo states with full-fledged judgmental content involving concepts like freedom or agency.

I do not mean to suggest that this distinction is a sharp one. It wouldn't surprise me if the two kinds of content blur into one another, via a spectrum of intervening types of psychological state and/or a spectrum of increasing forms of conceptual sophistication in different kinds of creatures. Also, it may well be that the two kinds of content can interpenetrate to a substantial extent, at least in creatures as sophisticated as humans. It is plausible, for instance, that humans can have presentational contents the possession of which require (at least causally) a fairly rich repertoire of background concepts that can figure in judgmental states. One can have presentational experiences, for instance, as-of computers, automobiles, airplanes, train stations – all of which presumably require a level of conceptual sophistication that far outstrips what dogs possess.

"divergence miracle" occurs shortly before the agent chooses/acts otherwise than how the agent chooses/acts in the actual world. I am even more unhappy with satisfaction conditions according to which some "accessible" possible worlds are allowed to differ somewhat from the actual world at all moments in time prior to the agent's non-actual choice/act. An idea that currently appeals to me is this: do the semantics of modals in terms of "scenario-specifications" that (a) are *epistemically* possible (relative to some contextually pertinent body of background information), and (b) need not be metaphysically possible. As regards modals about human agency, some such scenario-specifications will hold fixed the portion of the actual world that precedes a given agent's choice/act, will specify some way the agent chooses/acts that differs from the agent's actual-world choice/act, and will also specify that there are no violations of any actual-world laws of nature.

Briefly, the version of package-deal compatibilism that I favor comprises the following eleven theses. First, the presentational content of agentive phenomenology includes the aspect of self-as-source, which itself normally includes the aspect of core optionality (core "can/could do otherwise") as an essential component.[15] Second, the presentational intentional content of agentive phenomenology has satisfaction conditions that are compatible with state-causal determinism. Third, this compatibility is a non-manifest feature of agentive phenomenology; i.e., one cannot reliably tell, just on the basis of careful introspective attention to one's own agentive experience and the exercise of one's conceptual competence in judgment-formation, whether or not the compatibility hypothesis is true. Fourth, despite the compatibility of agentive phenomenology with state-causal determinism, a bodily event that is experienced as one's action cannot also be *experienced* as state-caused, either by non-mental states or by mental states. Fifth, the presentational aspect of core optionality remains present as an essential component of normal agentive phenomenology even when one experiences oneself as acting under coercion or duress. Sixth, an essential aspect of experiences of state-causation, including experiences of one's own bodily motions as state-caused, is the presentational aspect of *inevitability* – i.e., the aspect of inevitability *given the circumstances and the causing events*. Seventh, the two theses lately mentioned jointly explain the phenomenological mutual exclusion described in the fourth thesis: this exclusion results from the core optionality aspect of agentive phenomenology on one hand, and from the inevitability aspect of the

[15] Many recent versions of metaphysical compatibilism about free agency not only ignore free-agency phenomenology altogether (including the phenomenological aspect of core optionality), but also presuppose both (a) that the capacity to choose otherwise and do otherwise is incompatible with state-causal determinism, and (b) that the "can/could do otherwise" feature is simply never required for genuine free agency. Compatibilists who affirm claim (b) typically do so because of the conceivability of Frankfurt-style scenarios – and they then go on to affirm (a) by conceding to the incompatibilists the latter's own favored construal of 'can/could do otherwise'. All this seems to me to be seriously mistaken. Even if there are possible scenarios in which one exercises free agency even though it is not the case (because of a preempted potential cause waiting in the wings) that one can/could do otherwise, it doesn't *begin* to follow that the capacity to do otherwise is *never* required for genuine free agency. On the contrary, that capacity remains a *defeasibly* necessary condition for free agency, Frankfurt-style cases notwithstanding. My three biggest complaints about dominant versions of metaphysical compatibilism in the recent philosophical literature are (1) that they ignore free-agency phenomenology, (2) that they grossly overestimate the (quite limited) significance of Frankfurt-style scenarios, and (3) that they concede to incompatibilists the contention that if determinism is true then people can never choose or act differently than they actually do choose and act.

phenomenology of state-causation on the other hand. One cannot experience an item of one's own behavior both as inevitable and as something that one could have refrained from doing.

Eighth, at the level of *judgmental* intentional content, the concept of free agency involves a feature that is probably not exhibited by the free-agency aspect of *presentational* intentional content – viz., implicit contextual parameters that determine, in context-specific ways, contextually operative standards of satisfaction. For instance, in many contexts the standards operate in such a way that an action performed under extreme coercion – e.g., with a gun in one's face – do not count as free. I.e., under the contextually operative standards, the *judgment* that such an action is not free is correct. [16] (In other contexts, however, the concept of freedom is correctly used in such a way that its satisfaction conditions coincide with those for the core optionality aspect of sensory-experiential intentional content – for instance, when one says "I could have refused to give the gunman my wallet, although that would have been a foolhardy thing to do; thus, I exercised freedom of choice in giving it to him").

Ninth, the implicit contextual parameters governing the judgmental concept of free agency can take on a limit-case setting in certain contexts of judgment or conversation – i.e., a parameter-setting under which an item of behavior counts as a free action only if (i) it is not state-causally determined, and (ii) it comes about as a result of metaphysical-libertarian "agent causation" involving the self as a godlike unmoved mover.

Tenth, the satisfaction conditions for *presentational* free-agency intentional content – i.e., for free-agency *phenomenology* – coincide with certain non-limit-case, compatibilist, satisfaction conditions for *judgmental* free-agency intentional content. The satisfaction conditions for agentive phenomenology do *not* coincide with the incompatibilist satisfaction conditions that accrue to judgmental free-agency intentional content when the implicit parameters at work in the judgmental concept of free agency have extremal, limit-case, settings.

Eleventh, the metaphysics of free agency is constrained by the intentional content of free-agency phenomenology, and thus is also constrained by the (matching) intentional content of everyday, non-limit-case, ascriptions of free agency. So, since the phenomenological content and the conceptual content

[16] Such judgments will normally be keyed to certain aspects of phenomenology too, aspects that are superimposed upon the underlying phenomenology of core optionality – e.g., the phenomenology of duress under threat, the phenomenology of moral imperativeness, and the like.

are compatibilist, free agency itself is a phenomenon that is compatible with state-causal determinism.

Elsewhere, sometimes collaboratively, I have set forth arguments in support of the various theses constituting this version of package-deal compatibilism. Contextualist compatibilism about the judgmental concept of freedom, in a form that acknowledges limit-case parameter-settings that are incompatibilist, is defended in Horgan (1979), Graham and Horgan (1994), Henderson and Horgan (2000), and Horgan (forthcoming). Other aspects of the full package-deal are defended in Horgan (2007a, 2007b, in press-a, in press-b), and in Horgan and Timmons (in press). I will not argue for the position here, because of space limitations.

I do recognize that when one attends introspectively to one's free-agency phenomenology, with its presentational aspect of self-as-source which itself includes the aspect of freedom as an essential component, and when one simultaneously asks reflectively whether the veridicality of this phenomenology requires one to be an "agent cause" in the sense espoused by metaphysical libertarianism, one feels *some* tendency to judge that the answer to this question is Yes. If the position I have sketched is correct, then this tendency embodies a mistake: the satisfaction conditions of free-agency agentive phenomenology do not require heavyweight, metaphysical libertarian, "agent-causal freedom," and do not require the falsity of state-causal determinism. I certainly acknowledge that a theoretically adequate version of package-deal compatibilism should provide a plausible explanation of this mistaken judgment-tendency – an explanation of why the tendency arises so strongly and so naturally, once the compatibility issue is explicitly raised. I have addressed this challenge elsewhere, e.g., Horgan (2007a, 2007b, in press-a, in press-b). Although I lack the space here to summarize the "respectful debunking" explanation I have offered for incompatibilist judgment tendencies, let me just say that my proposed explanation draws on two principal resources: first, that fact, already stressed, that agentive phenomenology and the phenomenology of state-causation are mutually exclusionary, and second, the contextualist element that I claim is operative in judgmental attributions of free agency.

So the version of package deal compatibilism I favor, which is contextualist about the concept of free agency, allows for a fairly plausible explanation of the incompatibilist-leaning judgment-tendencies that naturally tend to arise when one asks whether free-agency phenomenology is compatible with state-causal

determinism. When one factors this into the mix, alongside the various convergent forms of largely abductive evidence (not set forth here) that favor both phenomenological compatibilism and conceptual compatibilism, I think a strong case can be made in support of an overall position that is phenomenologically compatibilist, conceptually compatibilist about everyday free-agency ascriptions, and metaphysically compatibilist.

5. Conclusion

Although the rich and distinctive phenomenology of agency went largely ignored in mainstream philosophy of mind in the twentieth century, it is now receiving renewed attention in that branch of philosophy. Agentive phenomenology also received far too little attention in twentieth-century philosophical discussions of freedom and determinism – with advocates of compatibilism probably being the worst offenders. It is time to bring the phenomenology of free agency explicitly into the freedom/determinism debate, and to accord it significant weight. A complete treatment of the freedom/determinism issue should address three topics together: the phenomenology of free agency, the concept of free agency, and the metaphysics of free agency. All else equal, a package-deal treatment of these topics should be strongly internally coherent – i.e., it should treat the phenomenology of free agency as strongly constraining both the concept of free agency and the metaphysics of freedom. This theoretical desideratum would spell big trouble for compatibilism if one could reliably ascertain, directly on the basis of introspection, that free-agency phenomenology has metaphysical-libertarian satisfaction conditions. But there are strong reasons to think that introspection is simply not that powerful – a fact that opens up room for abductive considerations to enter the dialectical mix. Once such considerations are properly brought to bear and given their due epistemic weight, I maintain, the overall package-deal position that will look best in terms of theoretical cost-benefit evaluation will be phenomenologically compatibilist, conceptually compatibilist (yet also conceptually contextualist), and metaphysically compatibilist.

REFERENCES

Chisholm, R. (1964). *Human freedom and the self. The Langley lecture* (University of Kansas). Reprinted in J. Feinberg & R. Shafer-Landau (Eds.) (2002), *Reason and responsibility: Readings in some basic problems of philosophy,* 11th Edition, (pp. 492-499). New York: Wadsworth.

Chisholm, R. (1995). Agents, causes, and events: The problem of free will. In T. O'Connor (Ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, (pp. 95-100). New York: Oxford University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*(3), 829-839.

Graham, G., & Horgan, T. (1994). Southern fundamentalism and the end of philosophy. *Philosophical Issues*, *5*, 219-247.

Henderson, D., & Horgan, T. (2000). What is a priori and what is it good for? *Southern Journal of Philosophy, Spindel Conference Supplement*, *38*, 51-86.

Horgan, T. (2007a). Agentive phenomenology and the limits of introspection. *Psyche*, *13*(2).

Horgan, T. (2007b). Mental causation and the agent-exclusion problem. *Erkenntnis*, *67*(2), 183-200.

Horgan, T. (in press-a). Causal compatibilism about agentive phenomenology. In T. Horgan, M. Sabates & D. Sosa (Eds.), *Supervenience and Mind: Essays in Honor of Jaegwon Kim*. Cambridge, MA: MIT Press.

Horgan, T. (in press-b). Introspection about phenomenal consciousness: Running the gamut from infallibility to impotence. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness*. Oxford: Oxford University Press.

Horgan, T., Tienson, J., & Graham, G. (2003). The phenomenology of first-person agency. In S. Walter and H. D. Heckmann (Eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, (pp. 323-340). Exeter: Imprint Academic.

Horgan, T., & Timmons, M. (in press). Introspection and the phenomenology of free will: Problems and prospects. *Journal of Consciousness Studies*.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies*, *11*(7-8), 162-179.

Stephens, G. L., & Graham, G. (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, MA: MIT Press.

The Decisions of Consciousness and the Consciousness of Decisions

Mauro Maldonato *
mauro.maldonato@unibas.it

ABSTRACT

For a long time the study of motor decision making has essentially been based on the mechanical neurophysiology of the connections between nervous structures. Empirical research and theoretical reflection have in this way been dominated by reflexological and cybernetic models without plausible alternatives. The tendency to separate the mental functions from the body, almost as though they were independent systems, has at times had negative consequences. Indeed, whether dealing with language or other cognitive and perceptive functions, the mind is profoundly influenced by the motor sphere, the oldest from an evolutionary point of view, which depends on the cortex, the basal ganglia and the cerebellum that contain motor, motivational and cognitive components. The ever-growing debate in the cognitive neurosciences, the philosophy of the mind and phenomenology shows that the time for a conceptual and epistemological change is growing nearer, a change which puts the idea of embodied consciousness and cognition back at the centre of the research being conducted.

1. The Matrix Controversies of the Motor Action Models

In the most famous of his *Croonian Lectures*, the English neurologist John Hughlings Jackson, father of modern neurology, noted:

> That activities of the highest, least organised, nervous arrangement, during which consciousness, or most vivid consciousness arises, are determined by activities of lower, more organised, nervous arrangements, I firmly believed. As I have said, in effect, states of consciousness attend survival of the fittest states

* University of Basilicata

> of centres representing all parts of the organism as one whole. Roughly speaking, the highest nervous states are determined from below, and not by autocratic faculties acting upon the highest part of the highest centre. (Jackson 1884, p. 706)

Despite the many, and often incorrect, interpretations of his philosophy, Jackson can be given the undeniable credit for having moved the neurophysiological debate of the 1800's from a relational life model founded on reflection (the automatic response that causes the simultaneity of forms and movements), to another model in which motor functions descend mechanically from cortical structures, which are the biological basis of rationality, imagination, logical thought and still more. For more than two centuries, reflex had been the dominant paradigm not only for philosophers such as Descartes, but also for the majority of neuroanatomists, neurophysiologists and neuropathologists. Jackson considered identifying the site of a lesion, a functional centre and anatomic location, to be erroneous, because ontogenesis realizes but above all directs the organism, integrating at a higher level that which is integrated at a lower level. At the centre of his research are the *functional metamorphoses*, whose temporality impresses rhythm and movement, guiding the relational life of every living being. Time, in fact, does not influence only the development of forms and movements, but also assigns a functional hierarchy to them.

The Jacksonian idea, according to which the evolution of the nervous system is characterized by ascending dynamics – from the more organized lower levels towards the less organized higher centres of the *highest level* (from the most automatic to the most voluntary) – introduced a new dimension into the debate of that age on the organization of the nervous system (1884). A concept that is so conditioned by the idea of evolution joins the notion of overlap to that of hierarchy, the notion of mechanisms to that of integration. In his vision the spatial (nervous) structures are subordinate to the flow of time: in this way, that which is lower (that is, more fixed) is subordinate to that which is higher (more mobile). The natural finalism in the hierarchy of nervous functions confers upon the concept of integration logical-sequential characteristics, according to which the lower or instrumental functions controlled by the *highest level* become subordinate like words to syntax, or means to an end (Ey 1947).

In spite of its apparent mechanism, the Jacksonian idea of an autonomous ontogenesis of relational life makes the principle of hierarchy

dynamic and, therefore, functional to that same process of integration: which is then nothing other than a sensorimotor coordination, a link between the present and the past, between imagined and perceived. In this sense, that which defines the *highest level* is its contingency (its freedom), and the same concepts of "automatic" and "voluntary" represent the levels of the functional hierarchy, whose morphology and nervous structures represent the free movement of relational life (Jackson 1932). The very notion of a "centre of consciousness" – the most controversial Jacksonian theme, considered by some to be the stumbling block of his hierarchical theory of functions – remains the most important issue in the current neuroscientific debate. Consciousness, the *highest level* of the evolution of the nervous system is, for Jackson, the structural-functional basis for the unfurling of the mind's activities: its very organization (Evans 1972). The order of consciousness is, in fact, sustained by multiple horizontal levels, every one of which is in a structural and functional *continuum* with various phenomenological occurrences (Maldonato 2009). It is such characteristic that allows for the integration processes of the activities of thought and of the programming of motor activities (even when only representational). Planning an action, in fact, always requires predicting its consequences, and this type of prediction is the result of model action activity. In this sense, thought and motility are tightly linked on both a phylogenetic level as well as an ontogenetic one. This link has over time produced an enormous archive of extraordinarily fluid motor repertories. The progressive refinement of the relation between the motor and the pre-motor cortex is at the origin not only of motor behaviours (such as the ability to construct and manipulate objects), but also of the acquisition of competences from structures such as Broca's area and the basal ganglia, which control the motor aspects of language. It must be said, however, that language is not an individual and autonomous system, but rather the product of a sophisticated coordination between systems and cerebral areas that are tied to the representation of objects, to perception and to the very motility of the body.

2. The Sense of Movement and Embodied Action

On a phylogenetic level sensory and motor activities – the basis of the development of various cognitive functions – have the longest history. The

wide range of structures in the human nervous system show, on the one hand, how complex the evolution of motor control has been and, on the other hand, its impact on other functions: from language to motility and so forth (Jeannerod1994). The motor and muscular systems are high-priority systems and their activation triggers the inhibition of the perceptive, sensory, attention, and other systems. This fact is even more readily apparent if one considers that, in animals, movements are linked to the carrying out of actions essential to survival, such as escape, attack, searching for food and the selection of a sexual partner. The activation of some muscles (even when only activated potentially, such as in the case of muscle tensing) involves the activation of other muscles, the reduction of sensations, the limitation of the flow of ideas and still more. This means that motility has not only direct motor consequences but also general effects on other systems. While it is true that movements depend largely on cerebral motor areas, it is in fact the whole nervous system that presides over the control of motility (MacKay 1987). The same cortical areas that decode sensations – through which we perceive muscle tension or the position of a limb – inform us retroactively about the execution of a particular movement. Without this function the movement would be imprecise, rough or completely blocked.

As it is known, muscles are controlled by the pyramidal neurons of the motor cortex, which are connected through the spinal marrow to motor neurons situated therein in order to reach, from there, the peripheral muscular fibres. Every muscular movement – such as moving a finger, shaking hands, crying and so on – involves the activation of the nervous-muscular neuron-fibre sequence. However, motor action is extraordinarily more complicated. In fact, if it is true that the composition and harmony of movements is guaranteed by the base ganglions and by the cerebellum – it is in these structures that the memory of the sequence of muscular actions are conserved, actions that allow us, for example, to centre in on a target with an arrow, pick a small flower, or dial a number – they constitute only the infrastructures of the movement: the planning and the execution of the movement depend, instead, on other cortical and subcortical structures (Adams *et al.* 2005).

Today, the relationship between the complexity of a motor action and the number and type of nervous structures involved is clearer. It has been observed, for example, that even simple and localized movements like the flexing or the stretching of the index finger of the right hand involve the activation of the primary motor area and of the somatosensory area of the

contralateral hemisphere. These are areas that are activated when a more complex movement is in action: for example, when subjects are asked to touch the tip of their right thumb to, in the following order, the tip of their index finger, middle finger, ring finger and pinkie finger of the same hand; although it must be said that in this case even the supplementary motor cortex and the prefrontal cortex are activated, the latter being activated even when the movement is simply imagined. In the case of the imagination and execution of a complex movement, the prefrontal area and the supplementary cortex are bilaterally activated, that is to say that there is activation even in the hemisphere not involved in the execution or imagination of the motor act (Brown and Marsden, 2001). This bilateral stimulus could correspond to the activation of an abstract plan of the movement or reflect a variety of motor plans oriented towards the same goal.

There are studies that indicate that it is first the prefrontal cortex (the decision to act) that is activated, then the supplementary cortex (involved in the plan of action) and, finally, the motor cortex, which implements and modulates the action based on the proprioceptive information that reaches the somatosensory cortex (Brown and Marsden 2001). Ultimately, the sequence of movements is due to two different circuits: an internal one, which involves the supplementary area, the basal ganglia and the temporal lobe, and takes over when a motor ability becomes habitual because it is guided by an internal representation of the action; and an external one, which includes the parietal lobe, the premotor area and the cerebellum, involved in direct movements or movements guided by spatial representations.

3. The Predictive Brain

The execution of remarkably complex actions, such as those of a musician at a piano, is much more articulated than what experiments on the planning and execution of simpler movements reveal, contextualized and guided as they are by the judgement of the performer. All of this was already clear to Lotze who, in the mid 1800's, wrote:

> We see in writing or piano-playing a great number of very complicated movements following quickly one upon the other, the instigative representations of which remained scarcely a second in consciousness, certainly not long enough to awaken any other volition than the general one of

> resigning one's self without reserve to the passing over of representation into action. All the acts of our daily life happen in this wise: Our standing up, walking, talking, all this never demands a distinct impulse of the will, but is adequately brought about by the pure flux of thought. (*quoted in*: James 1952, p. 791)

Beyond the musical interpretation and the talent of the single performer, all of the components of that musical ability derive from the complex interaction between motor learning, temporal processing and sequencing, in which a crucial role is played by the relations between the cortex and the basal ganglia. In reality, the line between perception and action is not well-drawn as one might believe when basing oneself on the description of the execution of a motor task (Berthoz 1993). If the cerebral structures' capacity for processing is considered, rather than their specific function in the execution of a task, not only does the crucial role played by the parietal lobe in the perception and execution of an action become clear, but also that of the basal ganglia in the sequencing of movements, language or ideation. Despite being parts of different systems, perception and action constitute integrated functions. In light of these considerations, subordinating motor functions to higher cognitive activities and classifying the body as an inferior entity to that of the mind appears implausible. The body and its movements are at the origin of the abstract behaviours of which we are proud, beginning with language which gives form to our mind. For example, the evolution of some motor behaviour, such as the ability to construct and manipulate objects, selected an order of movements based on a sequence of cause-effect links. This led the motor and premotor cortex to develop a growing ability to generate interlinking movements, inducing even *Broca's area* to produce the verbal gestures and the sequences of syllables that are at the basis of communication. In this sense, pronouncing a sequence of syllables is like sculpting bronze or sharpening a blade: this control of motility preceded language, but also contributed to structuring it as an internal motor logic (Oliverio 2009).

It is rather probable that the logic of the body and of its movements constituted the foundation on which, over time, the operational logic of language structured itself. In terms of physical experiences many motor operations have been so important that they have progressively supplied the infrastructures for the development of symbols and metaphors used in language, translating themselves over time into classes of perceptions, behaviours and universal linguistic conventions (Lakoff and Johnson 1980).

4. Embodied Memories, Goals and Plans of Action

Perception is, by its very nature, multisensory. It uses multiple reference systems adapted to the actions in progress. In fact, while receptors measure derivatives, the brain mobilizes a repertory of prototypes of forms, faces, objects, and even synergies of movements. During its progress, evolution selected simplifying laws in the geometric, kinematic and dynamic properties of natural movements. But perception is also predictive, thanks above all to memory, which uses the consequences of past actions in order to predict those of future actions (Berthoz 1998). Whether shaking hands, writing a letter or performing another action, every executive act requires a behaviour directed towards a goal, a behaviour made possible thanks to the control of a series of nervous structures and mental processes that process information.

Because of its complex relations with the other cortical areas and subcortical nuclei, the frontal cortex is at the centre of the executive functions: from the memory of work (which allows one to remember the beginning of a sentence once completed) to the behaviour directed towards a goal (which implies a continuous re-modulation of information with the passage from one plan of action to another and the continuous verification of the consequences of our actions). Such functions depend on the prefrontal cortex (in human beings it accounts for approximately half of the frontal lobe), which being linked to all of the other cortical areas and to a large part of the subcortical structures is directly or indirectly involved in all of the executive functions (Miller *et al.* 2002).

But how do we succeed in formulating plans of action corresponding to specific goals? A plan of action involves a hierarchy of relevant actions and irrelevant actions. In addition, it can be part of a vast plan consisting of immediate objectives or of sub-plans matching the principal objective. These complex functions involve the planning and the choice of an action, the monitoring of its execution, and the reinforcement tied to the reaching of the desired goal.

Since the by now classic studies of Leonardo Bianchi (1889) on the effects of bilateral ablation of the prefrontal cortex of primates, the executive functions of the motor system have been attributed to the prefrontal lobes. In order to fully grasp the subtle and complex relations of the prefrontal cortex with behaviour it is useful to understand the distinction between the lateral prefrontal cortex and the medial prefrontal cortex. The lateral prefrontal cortex

can be further subdivided into the dorsolateral prefrontal cortex (which selects the information) and the ventrolateral prefrontal cortex (which stores the information). The medial prefrontal cortex can also be subdivided into two important areas: the anterior cingulate cortex (which identifies the errors of specific behaviour) and the superior frontal gyrus which seems to be involved in the selection and the execution of a task (Rushworth *et al.* 2004). In reality, these anatomic-functional subdivisions and their implications on behaviour are not always so clear-cut. In fact, between anatomic areas and functions it is not infrequent that overlapping levels are observed, a fact that encourages researchers to be very careful when defining the role of different frontal and prefrontal areas.

This intricate neuronal geography propels us to reconsider the integration processes between frontal and prefrontal areas, whose collaboration creates that complex phenomenon called motor control, the dynamics of which are in some ways the opposite of those of perception. Indeed, if perceiving means putting the external world into an image, acting means representing to oneself the desired consequences of a movement which is being carried out while it is being carried out. In this sense, the execution of a movement has to do with a representation of the environment, beginning with the information made available by the parietal cortex and by the hippocampus which, as is known, is a structure involved in numerous aspects of spatial memory (Oliverio 2008).This information passes to the premotor cortex which, so to say, 'projects' the movement and, finally, to the motor cortex which carries out the action.

As we have seen, motor control and its execution depend on cortical and subcortical structures, among which we find the basal ganglia that play a fundamental role in the control of spatial memories, of motor actions in a specific context and of the motivational components of learning. In this schema, the cortex and the basal ganglia plan the action, the execution of the movement and the control over its state of execution, in close collaboration with the cerebellum, the red nucleus, the striated muscle and other subcortical structures. For almost a century and a half, motor functions were instead considered to be directly dependent on superordinate structures, such as cortical ones, considered to be the basis of higher cognitive activities: rationality, creativity, and thought. In reality, thought activities and motor activities (even when only representational) are always closely correlated. Whether imagining, planning or acting, it is always the same area of the brain

that is activated. The planning of an action always, in fact, requires the prediction of its consequences, and this type of prediction is the result of model action activity (Oliverio 2008).

The tendency to separate mental functions from the body has negative consequences. Whether dealing with language or other cognitive and perceptive functions, the mind is profoundly influenced by the motor sphere, which in turn depends on older structures such as the cortex, the basal ganglia and the cerebellum. The prevalence of a hierarchically superordinate vision of the mind (to the detriment of the motor sphere) has depended on true and proper philosophical misunderstandings, which are worth examining briefly. In contrast with the arguments that identify him as the greatest driving force behind modern philosophical dualism, Descartes shed light on the intimate and immediate relationship between mind and body. In the sixth of the *Meditations on First Philosophy*, the French philosopher argues that nature teaches him

> [...] through these very feelings of pain, hunger, thirst, and so forth, that I am not present in my body only as a pilot is present in a ship, but that I am very closely conjoined to it and, so to speak, fused with it, so as to form a single entity with it. For otherwise, when the body is injured, I, who am nothing other than a thinking thing, would not feel pain as a result, but would perceive the injury purely intellectually, as the pilot perceives by sight any damage occurring to his ship; and when the body lacks food or drink, I would understand this explicitly, instead of having confused feelings of hunger and thirst. (Descartes 2008, p. 57)

Descartes affirms that we are joined to our body, that the mind is mixed with the body as though it were one entity and that we are conscious of what happens in our body, although in a different way from how we are conscious of objects external to the body. In short, we do not look at our body as we look at other things. We do not have to check, for example, the position of our legs or whether we have our hands in our pockets. We know this information without having to verify it. Unlike those patients who, because of a vascular accident or another cerebral lesion, have lost the sense of the body's movement and of their own position in the space around them. As is known, in order to be aware of movement and of their own position these patients have to check the position of their own body, just as the Cartesian "pilot" looks at his own ship.

Beyond the necessary rereading of Cartesian philosophy, in evolutionary terms the human nervous system developed mainly in order to coordinate

perception and body movements and to increase efficiency in activities essential for survival such as hunting, coupling and raising offspring. As paradoxical as it may seem, evolution has favoured the development of knowledge for efficient action, not so much for reflection. James asks himself whether the simple idea of the effects of a movement is a sufficient motor stimulus or whether there is an additional mental antecedent, such as a decision or some other analogous phenomenon, in order to which there may be movement (James 1952). He advances the idea that a movement is always associated with a representation of its consequences and that every representation of a movement reawakens with the maximum level of intensity the real movement, every time it is not impeded by an antagonistic idea simultaneously present in the mind (James 1952). Following along the lines of Lotze, who believed that the imagination of a movement activated the same structures involved in its execution, James suggests that consciousness is always the consciousness of an action.

5. Decisions of Consciousness

During its different evolutionary stages biological life on our planet produced two main adaptations: to begin with it imprinted elements into the genetic code that would facilitate the periodic variability to environmental changes such as light, temperature, precipitation and still others; and secondly it equipped the animal nervous system with structures that would guarantee the sensory and motor activities developed through time (Maldonato and Dell'Orco 2010). Compared with higher animals human beings also have an internal representation of time, and this originates in the birth of conscious experience. It is through the conscious perception of time that, over the course of evolution, human beings have been able to achieve enormous adaptive and reproductive advantages.

As a neurobiological phenomenon distinct from awareness, consciousness originates in the cortical-subcortical space, even if it is only in the cerebral cortex that the experience of time is realized, that is, the unmistakable individual impression of continuous past experiences that is bound together with future expectations. And it is always in the cortex that the unification of time takes place, realized through the combination between nervous circuits and our conscious experience, to which we can add through introspection and

accounts in the 'third person'. Although it is an essential characteristic of consciousness, we know little about time. These notions revolve around the categories of succession and duration (Fraisse 1987). Succession implies the eminently cognitive distinction between the simultaneity and the sequence of a number of events – although not in an absolute sense, because when temporal scales of tens of milliseconds are used the reliability of our judgement becomes more uncertain. Duration instead implies the ability to understand sequential perceptive events as though they were simultaneous, that is to 'feel' the interval of time without discontinuity. In *Time and Free Will: An Essay on the Immediate Data of Consciousness* (1910), Bergson problematizes the spatialized vision of duration of the positive sciences by identifying two dimensions of conscious life: a superficial I, which is built on cognitive issues; and a fundamental I, which is built through the synthesis of consciousness. Before Bergson, it was the Eleatic philosophers and later Saint Augustine (*The Confessions*) who shed light on the problematic nature of the concept of the Present and who questioned time as the succession of present moments. How short can a moment be, that changing interval that flows from the past to the future and vice versa? According to James (1952) our consciousness of time originates in different speeds, which depend on the number of events or changes that we experience in a certain interval (neuroscientists would speak of a minimum necessary time for the emergence of neural events correlated to a cognitive event). This immaterial structure has been interpreted as the phenomenon of surfaces of a neural integration at wide range, tied to a diffuse synchrony: this being an interpretation that could clarify, through a dynamic reconstruction, both the invariant nature of events and the synchronization process of tangible experience (Petitot *et al.* 1999).

In reality, there is no agreement on the nature of the processes at the basis of succession and duration. In general, the most accredited hypothesis is that the perception of time takes place around the following orders of magnitude: below one hundred milliseconds it is possible to distinguish the beginning and the end of an event, its instantaneity; past five seconds the perception of the duration seems to be cut in half by memory (Fraisse 1987). The 'moments' of this *deceptive present* are believed to oscillate between 100 milliseconds and 5 seconds. Other hypotheses indicate that at the foundation of consciousness is a mechanism of temporal unification of neuronal activities that synchronizes impulses in medium oscillations of 40 Hz (Crick 1994). These oscillations are not believed to codify additional information, but they are thought to unify part

of the existing information in a coherent perception. Our consciousness, therefore, would not be generated by the action of a specific zone of the brain, but by the concomitant activation of a series of neurons distributed in the brain. Such oscillations are a necessary but insufficient condition for the production of conscious experience.

The phenomena of general neuronal activity as seen by EEG originate in the activation, parallel inhibition and synchronization of multiple neuronal circuits. This is a dynamic balance, in which every event, lasting from 100 to 200 milliseconds, reflects the activation of a distributed and parallel neural network that is translated into the contents of consciousness, such as an abstract thought or a visual image (Le Van Quyen *et al.* 1997). In certain conditions, there are areas in which neuronal oscillations play a crucial role. In addition, certain states of consciousness (alertness, falling asleep, waking, etc.) and pathologies such as depression, epilepsy, and Parkinson's disease cause different registrations of thalamic-cortical rhythms (Charney *et al.* 1996), whose duration varies with the variation of clinical populations. For example, in paranoid schizophrenics they are shorter (Torrey *et al.* 1994), whereas in manic patients the rhythms show continuous changes (Goodwin and Jamison 1990) and so on. It is not implausible to maintain that these neuronal harmonies and discords give way to the emerging phenomena that make subjective experience possible. A thus-constructed model would allow us to do without metaphysical entities such as the *central theatre* of Baars (1997), the *homunculus* of Dennett (2005) or any other metaphysical entity, letting the I of neuronal organization emerge and, therefore, the subjectivity of the physical brain. Careful reflection on the concept of temporality encourages the reconsideration of some aspects of consciousness that seem obvious. The first aspect to be reconsidered is the unity of conscious experience, which disappears as soon as it is considered on the basis of time scales of milliseconds (Roeckelein 2000); the second is immediacy, a phenomenon sometimes too quickly attributed to consciousness. We have already seen previously how continuous visual information is connected to different processes that require certain intervals of time. Furthermore, the milliseconds relating to the duration of these processes are irrelevant (Richelle *et al.* 1985) and no piece of information can reach consciousness until at least half a second has passed after its arrival in the cerebral cortex.

In reality, experimental research has yet to propose convincing solutions for the problem of the experience of time. This is perhaps because this

disconcerting enigma is different from the one relating to the cerebral areas and structures that are at the origin of phenomena and experiences, which can be studied today through *brain imaging* methods (Posner and Raichle 1994, Zeman 2001). As the origin and structure of consciousness, temporality joins together the different levels of neurophysiological and phenomenological reflection. An efficient research method is composed of cerebral activation studies (PET, fMRI, MEG, event-related potentials) which allow for the exploration of the central nervous system before and after an adequate stimulus: the presentation of ambiguous visual stimuli, the transition from general anaesthesia to reawakening, the passage from a vegetative state to a minimally conscious one and still others. For example, the rekindling of the activity of the re-entering thalamic-cortical circuits, in a patient who was first 'vegetative' and then 'minimally conscious', shows the importance of the role of the connections between the intralaminar nuclei of the thalamus and the frontal and parietal associative cortices in the maintaining of consciousness. Here, a fundamental task is performed by the *Ascending Reticular Activating System* (ARAS) – a system composed of the reticular formation, the thalamus and the thalamic-cortical projection system – which presides over the diffuse activation of the cerebral cortex in states of wakefulness and alertness, states necessary for the formulation of the contents of consciousness (Moruzzi and Magoun 1949). This is a distributed system, not circumscribable to the reticular nuclei of the encephalic trunk (Plum and Posner 2000) that projects itself in a descending direction towards the spinal cord and, in an ascending direction, towards the cerebral hemispheres. Each one of its constituent nuclei has particular anatomic, physiological and biochemical characteristics: those that modulate the functioning of the cortex reside in the upper two thirds of the pontine tegmentum, others in the lower third of the pons and in the bulb – that is why, in stroke patients, isolated lesions of the pons can cause a coma even in the absence of mesencephalic damages (Wilkinson and Lennox 2007). It is not without significance, moreover, that some nuclei of the cerebral trunk surpass the thalamus in order to connect directly with the frontal-basal cortex, from which the bilateral projections diffused to the cerebral cortex originate; or that other nuclei go beyond both the thalamus and the frontal-basal cortex to reach wide areas of the cerebral cortex; or that, finally, other nuclei are connected with the reticular nucleus of the thalamus and not with the intralaminar nuclei.

This unique neuronal geography allows us to consider the functions of the ARAS as being much more wide-ranging and complex than those linked to the

simple 'desynchronization' of the cerebral cortex (Mancia 1994), also essential to the state of wakefulness and attention. Then there are the non-specific thalamic-cortical projections, such as the activation of the thalamic-cortical circuit at a high oscillatory frequency, projections fundamental to the essential functions of consciousness. Studies on cerebral activation (Laureys *et al.* 2004) have demonstrated that, in patients in a vegetative state (a state of wakefulness without content), the connectivity between cerebral areas that are normally connected is lost: in particular, between the primary cortical areas and the associative multimodal ones (the prefrontal, premotor, and parietal-temporal areas, the cortex of the posterior and precuneous gyrus cingulate) or between these cortical areas and the thalami. This leads one to wonder whether the exclusive role of ARAS in determining consciousness should not be reconsidered, rethinking consciousness as the effect of the interaction of an enormous variety of *qualia* and of distinct perceptions implied in the distributed and dynamic activity of the thalamic-cortical nucleus.

In general, consciousness is a stable and at the same time variable temporal event generated by an interaction of different levels – neural infrastructures, qualitative-subjective experiences and functional units – that are logically interrelated. This is a structure-function that is radically different from the other phenomena of the natural world (Maldonato 2007), one that emerges through an order in which the schema produced by the system's elements cannot be explained by the individual action of the system's single constituents, but rather by the synergy between its elements: this being a phenomenon that can be found both in elementary environments and in extremely complex ones.

There now seems to be a general consensus that at the basis of consciousness there is synchronization between different cerebral regions, and that this form of temporalization constitutes a deciding factor in the integration processes of neuronal information. However, the question remains open as to the nature of the passage from the neuronal level to that of perception and, finally, consciousness. It is not enough, in fact, to postulate an explanatory principle (chronological time or any other synchronizing function) without taking the mechanisms for accomplishment into account. Varela (1996) has long insisted on the necessity of considering consciousness as an emerging phenomenon, in which local events can give rise to properties or global objects in a reciprocal causal co-involvement. These are structural invariants incompatible with the continuous representation of linear time inherited from

classical physics (Prigogine 1986, 1997). More recent theories on consciousness hypothesize a minimum necessary amount of time for the emergence of neural events that connect themselves to a cognitive event (Dennett and Kinsbourne 1992). This temporality can plausibly be attributed to long-range cerebral integration linked to diffuse synchrony: an event that would shed light on phenomenological invariants, restoring tangible experiential content to the synchronization process.

For a long time scholars focused on the concept of the unitarity and the permanence of consciousness in time. Today, instead, numerous studies show that consciousness is a plural process that encompasses different contents in itself simultaneously, each element of which has its own intentionality (Zeki 2003, O'Brien and Opie 2000).

But what are the biophysical mechanisms of the unified experience of consciousness? And how does this internal plurality unify the different contents? There seem to be two possible models. The first model hypothesizes that consciousness is generated by a central neural system, in which duly integrated information is first represented and then brought to consciousness. In this schema consciousness appears to be the result of the work of the central neural system that generates different contents and representations, a phenomenon taking place exclusively in the brain. In the second model the simultaneous co-activation of the contents generated by distributed structures in the brain are believed to give rise, ultimately, to the phenomenon of consciousness. Consciousness would in this way be generated by distributed cerebral mechanisms – both cortical and subcortical – the contents of which, each element being independent one from the other, are exposed to intrasensory and intersensory (environmental) influences. The contents of the distributed cerebral mechanisms and the intrasensory and intersensory influences affect each other reciprocally and thus co-determine conscious experience. It is in this fine line that the distinction between a unitary model and a plural model of consciousness lies.

Ramachandran (2004) has a number of times discussed the plausibility of a model that integrates visual, auditory, tactile and proprioceptive experiences as well as other experiences. These individual spheres, in a relatively independent way, can be altered or neutralized without influencing the other spheres. Experimental evidence relating to the consequences of lesions and ablation of cerebral areas show that if, on the one hand, it is possible to lose the capacity to visually grasp movement, conserving however the other aspects of visual

experience (Zeki and Bartels 1998), on the other hand, it is possible to lose the sensation of colour, without losing visual experience and the experience of movement. Studies on the deficits caused by lesions on the level and kind of functional specialization and cerebral localization have shown that the brain works on a large scale, between procedures and domains that are reflected in specific anatomical districts (primary visual processing in the occipital cortex, auditory processing in the temporal cortex, planning and memory processing in the frontal cortex), while specific functions are realized in well-demarcated anatomical districts and locations (for example, the visual motor function takes place in area V5 and that of colour in V4). The zones of the brain that program particular informational content are those in which the contents come into consciousness. For example, different events from a visual scene, presented simultaneously, are not perceived with the same duration. This multiple asynchrony seems to prove that consciousness is the integrated result of countless micro events more than a unitary faculty (Zeki 2003).

But how can these multiple neural events restore to us the impression of a unitary subjectivity? And which paths lead to the composition of the Self and of consciousness? Concepts such as 'unitary subjectivity' and the 'Self' remain problematic. Here, we will limit ourselves to affirming that the Self emerges when the individual events produced by the brain are sufficiently representational, coherent and close-knit. In the absence of neurological and psychiatric disorders, we experience a structured world of distinct objects ordered in space, organized according to regularities and contents within meaningful spatial-temporal schemas: extramodal contents (colours, forms, etc.) and intramodal contents (proprioceptive, auditory and visual). In reality representational cohesion is not an invariant characteristic of conscious experience, but the result of a selection through which the brain searches for the path of its own integration. Ultimately, the Self has to do with a regulatory activity of consciousness that processes and maintains such plurality in an interweaving of local contents in contact with each other. In such a model, consciousness appears not as a hierarchical entity, but as a multiple horizontal entity, whose representational cohesion is carried out by thalamic-cortical and cortico-cortical distributed circuits. All conscious experiences, beginning with those that are qualitative (*qualia*), become unified within the field of consciousness. In this sense, unity is implicit in qualitative subjectivity. But if our consciousness is determined by the play between these innumerable dynamics, then there are not only different conscious states unified in

subjectivity, but also aggregate underlying fields of consciousness. In other words, the unitarity of consciousness follows subjectivity and quality because there is no way to have subjectivity and quality without unity.

The issue of conscious subjectivity goes beyond the search for its neuronal correlates and even beyond the conceptual contraposition between consciousness and the unconscious. For example, in the phenomenon of vision the methodologically relevant question certainly concerns the neural coordinates of consciousness, but above all it regards the way in which visual experiences enter and become part of the conscious sphere. If the infrastructure behind the field of consciousness is the thalamic-cortical system – which reprocesses the information originating from the different districts in various sensory forms (visual, tactile, auditory and so on) – from its operational neural levels one could remount to the structure of visual consciousness, of *qualia*, of temporal experience and still more. Nevertheless, the brain cannot generate conscious experience on its own: it is, in fact, only a necessary condition so that countless neuronal micro events may generate conscious perceptions of the world's objects (Varela *et al.* 1992). In this sense, an in-depth study of consciousness requires multi-level explanatory criteria: a *quantitative-categorial* criterion (attention, alertness, sleep, and coma); a *qualitative-dimensional* criterion (subjective experiences such as sensations, thoughts, and emotions); and a final criterion for the analysis of the different synchronic (the field of consciousness) and diachronic (the I and personality) types and levels of consciousness. At the present day, almost no one among scholars maintains that consciousness is characterized by a strict alternation between states of wakefulness and sleep. The constant variability of consciousness is demonstrated by numerous situations: from the clear and ready alertness of an airplane pilot to the attention levels of a student immersed in speculation; from the concentration of a monk in contemplation to the labile alertness of a drowsy or distracted individual. Something analogous can be said of sleep, which through the study of EEG correlates can be analyzed according to different qualitative and quantitative criteria (Mancia 1994). It must be noted, furthermore, that levels of consciousness are conditioned not only by physiological variations of the sleep-wakefulness rhythm, but also by the ingestion of anaesthetic drugs (which reduce the level of consciousness) or psychoactive substances (which increase attention levels).

Studies conducted on experimental animal models have shown that among the cerebral structures involved in the modulation of alertness are the *locus*

coeruleus (with adrenergic projection), the posterior portion of the hypothalamus (with histaminergic projection), other brainstem nuclei (with serotonergic and dopaminergic projection) and, above all, the intralaminar nuclei of the thalamus. The latter, in particular, play the essential role of synaptic relay for the diffuse cortical paths that regulate the synchronization of the cortical electrical activity registered by EEG. A lesion of these centres can cause a coma and vegetative states measurable using criteria such as those of the *Glasgow Coma Scale* (Teasdale and Jennett 1974). Expressions such as a loss of consciousness, a reduction of the level of consciousness, regaining consciousness, and others refer to this meaning of the term, essentially overlapping with the concept of *awareness*.

6. Conclusions

In this essay, it has been shown how numerous aspects of motor planning and of the intentional perception of an agent do not appear on the conscious level. The integration between these levels has a concrete meaning, which has effects on those conceptions of the mind that have been at the centre of the philosophical debates on the philosophy of action. Varela (1996) highlighted the role played by the body on the dynamics of perception; however, his reflection is still "disembodied", that is without empirical support. According to Berthoz (1998), the body is not only a thing, a potential scientific object of study, but also the necessary condition of experience. It constitutes the perceptive opening to the world: the primacy of perception is a primacy of experience, when perception reassumes an active and constitutive role and can be at the basis of action.

In the embryonic, fetal and infancy stages, action precedes sensation and not the opposite: first reflex movements are carried out and after they are perceived. We are normally led to emphasize sensations and perception, and particularly to retain that movement is essentially dependent on them. On the contrary, we could represent this sequence inversely through a schema in which one begins with movement in order to then consider the consequences that this has on the surrounding environment, namely the perception of the consequences and the modifications that this has on subsequent movements.

REFERENCES

Adams, R. D., Victor, M., & Ropper, A. H. (2005). *Principles of Neurology*. New York: Mc Graw-Hill.

Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.

Bergson, H. (1910). *Time and Free Will: An Essay on the Immediate Data of Consciousness.* London: George Allen and Unwin.

Berthoz, A. (1993). *Multisensory Control of Movement*. Oxford: Oxford University Press.

Berthoz, A. (1998). *Il senso del movimento*. Milano: McGraw-Hill Companies.

Bianchi, L. (1889). Semiotica delle malattie del Sistema Nervoso. Milano: Vallardi.

Brown, P., & Marsden, J. F. (2001). Cortical network resonance and motor activity in humans. *Neuroscientist, 7*(6), 518-527.

Charney, D. S., Woods, S. W., Krystal, H. H., & Heninger, G. R. (1996). Neurobiological mechanism of human anxiety. In B. S. Fogel, R. B. Shiffer & S. M. Rao (Eds), *Neuropsychiatry*, (pp. 257-286). Baltimore: Williams & Wilkins.

Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul.* New York: Scribner.

Dennett, D. (2005). *Sweet Dreams. Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press.

Dennett, D., & Kinsbourne, M. (1992). Time and the observer: The where and the when of the consciousness in the brain. *Behavioral and Brain Sciences, 15*, 183-247.

Descartes, R. (2008). *Meditations on First Philosophy with Selections from the Objections and Replies*. New York: Oxford University Press.

Evans, P. (1972). Henry Ey's concepts of the organization of consciousness and its disorganization. An extension of Jacksonian theory. *Brain, 95*(2), 413-440.

Ey, H. (1947). Système Nerveux et Troubles Nerveux. *Evol. Psych., XII*(1), 71-104.

Fraisse, P. (1987). Temporal structuration of cognitive processes: Discussion. In M. Olivetti Belardinelli (Ed.), *Comunicazioni scientifiche di psicologia generale, 15*, (pp. 26-33). Roma: Bulzoni.

Goodwin, F. K., & Jamison, K. R. (1990). *Maniac Depressive Disorder*. Oxford: Oxford University Press.

Jackson, J. H. (1884). The Croonian lectures on evolution and dissolution of the nervous system. *The British Medical Journal, 1*(1214), 660-663.

Jackson, J. H. (1932). *Selected Writings*, 2 Vols. London: Hodder Stoughton.

James, W. (1952). *The Principles of Psychology.* Chicago: Encyclopedia Britannica Inc.

Jeannerod, M. (1994). The representing brains: Neural correlates of motor intention and imagery. *Behavioural Brain Sciences, 17*, 187-202.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.

Laureys, S., Owen A. M., & Schiff N. D. (2004). Brain function in coma, vegetative state, and related disorders. *Lancet Neurology, 3*(9), 537-546.

Le Van Quyen, M., Adam, C., Lachaux, J. P., Martinerie, J., Baulac, M., Renault, B., & Varela, F. J. (1997). Temporal patterns in human epileptic activity are modulated by perceptual discriminations. *NeuroReport, 8*(7), 1703-1710.

MacKay, D. G. (1987). *The Organition of Perception and Action*. New York: Springer Verlag.

Maldonato, M. (2007). La coscienza prismatica. Un mosaico di forme incostanti. In M. Maldonato (Ed.), *La coscienza. Come la biologia inventa la cultura*, (pp. 13-60). Napoli: Guida.

Maldonato, M. (2009). Il meraviglioso algoritmo. Sulla struttura immateriale della coscienza. In M. Maldonato & R. Pietrobon (Eds.), *Pensare la scienza*, (pp. 41-68). Milano: Bruno Mondadori.

Maldonato, M., & Dell'Orco, S. (2010). *Psicologia della decisione*. Milano: Bruno Mondadori.

Mancia, M. (1994). *Neurofisiologia*. Milano: Cortina.

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* 357(1424), 1123-1136.

Moruzzi, G., & Magoun, H. W. (1949). Brain stem reticular formation and activation of the EEG. *Electroencephalogr. Clin. Neuro., 1*, 455-473.

O'Brien, G., & Opie, J. (2000). Disunity defended: A reply to Bayne. *Australasian Journal of Philosophy, 78*(2), 255-263.

Oliverio, A. (2008). *Geografia della mente*. Milano: Raffaele Cortina.

Oliverio, A. (2009). *La vita nascosta del cervello*. Milano: Giunti Editore.

Petitot, J., Varela, F., Pachoud, B., & Roy, J.-M. (Eds.) (1999). *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*. Standford, CA: Stanford University Press.

Plum, F., & Posner, J. B., (2000). *Stupor e coma*. Roma: SEU.

Posner, M. I., & Raichle, M. E. (1994). *Images of Mind*. New York: Scientific American Library.

Prigogine, I. (1986). *Dall'essere al divenire: tempo e complessità nelle scienze fisiche*. Torino: Einaudi.

Prigogine, I. (1997). *La fine delle certezze: il tempo, il caos e le leggi di natura*. Torino: Boringhieri.

Ramachandran, V. (2004). *The Emerging Mind*. London: Profile Books.

Richelle, H., Lejeune, J. J., Perikel, & Fery, P. (1985). From biotemporality to nootemporality: Toward an integrative and comparative view of time in behavior. In J. A. Michon & J. L. Jackson (Eds.), *Time, Mind and Behavior*, (pp. 75-99). Berlin: Springer-Verlag.

Roeckelein, J. E. (2000). *The Concept of Time in Psychology*. Westport: Greenwood Press.

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences, 8*(9), 410-417.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet, 2*(7872), 81-84.

Fuller Torrey, E., Bowler, A. E., Taylor, E.H., Gottesman, I. I. (1994). *Schizophrenia and Manic-Depressive Disorder: The Biological Roots of Mental Illness as Revealed by the Landmark Study of Identical Twins*. New York: Basic Books.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330-350.

Varela, F., Thompson, E., & Rosch, E. (1992). *La via di mezzo della conoscenza. Le scienze cognitive alla prova dell'esperienza.* Milano: Feltrinelli.

Wilkinson, I., & Lennox, G. (2007). *Manuale di neurologia*. Torino: Minerva Medica.

Zeki, S. (2003). *La visione dall'interno. Arte e cervello*. Torino: Bollati Boringhieri.

Zeki, S., & Bartels, A. (1998). Toward a theory of visuale consciousness. *Consciousness and Cognition, 8*(2), 225-259.

Zeman, A. (2001). Consciousness. *Brain, 124*, 1263-1289.

Epistemic Trust.
Outline for a Phenomenology of Shared Intentionality

Roberta De Monticelli *
demonticelli.roberta@hsr.it

ABSTRACT

Phenomenology is a method for thinking the (ontological) novelty of things, as irreducible to their (physical, biological, psychological) foundations. In this paper I shall exemplify this claim by addressing a question debated in contemporary philosophy of mind, analytical ontology, moral and natural philosophy, namely: what makes a human person out of a member of the biological species homo sapiens? A set of socially transmitted rules, a second cultural nature, seems to be a necessary condition for what we called primary self-constitution, the emergence of a "normally" behaving human subject. Epistemic trust is the basic condition for this transmission. The arguments for my claim are part of a general theory of acts, including voluntary actions, mental acts, speech acts and social acts, providing the foundation for a theory of personal identity and research in the field of social cognition.

Trust is a very intriguing subject for a phenomenologist. For phenomenology itself can be defined as a way of thinking based on the exercise of trust – albeit a peculiar kind of trust, that I'll term *epistemic trust.*

1. Epistemic Trust and the Culture of Suspicion

Phenomenology has been here for a century, and yet very few people do really understand its novelty. Too many thinkers or just scholars have usurped its beautiful name, without sharing in the least its spirit, without applying or

* Vita-Salute San Raffaele University

developing the methods for philosophical *research* on vital topics in our contemporary world, for which it had been devised.

What is, in fact, the spirit of phenomenology? I'll try to summarize it by this very notion of epistemic trust. I'll define epistemic trust as the systematic adoption of following key-principle: (ET) *Nothing appears in vain* (without a foundation in reality) – of course the reverse is not true: There is much more to discover in reality than what appears (otherwise no *research* would be needed, and we would be omniscient).

Epistemic trust is a style of thinking**,** which might be clarified through some more definite methodological principles. In this presentation I do not want to get into methodological details, though. The first thing I want to convey by this formula is that phenomenology has been so widely misunderstood, because we have not yet – not in the least – understood the whole depth of Plato's summons: *sozein ta fainomena,* to "save" phenomena. That is, things which are seen, things which appear*, fainomena* indeed.

Phenomenology so characterized seems to radically escape what the French philosopher Paul Ricoeur termed the "culture of suspicion". Under such a phrase I understand the mental attitude quite opposed to epistemic trust: a complete lack of confidence in the world of phenomena, that is in the ordinary world of our daily experience. This is both faithful and unfaithful to Ricoeur's own understanding of his phrase.

Faithful, on one hand. In his highly influential work, *Freud and Philosophy*, Ricoeur (1970) draws attention to three key intellectual figures of the twentieth century who, in their different ways, sought to unmask, demystify, and expose the real from the apparent; «Three masters, seemingly mutually exclusive, dominate the school of suspicion: Marx, Nietszche, and Freud» (Ricoeur 1970, p. 32).

On the other hand, Ricoeur's analysis focuses on a supposed false consciousness haunting – according to the three masters – a particular kind of experience – namely, religious experience. Religion is not about what it seems to be about. According to Marx, while religion appeared to be concerned with the lofty issues of transcendence and personal salvation, in reality its true function was to provide a "flight from the reality of inhuman working conditions" and to make "the misery of life more endurable". Religion in this way served as "the opium of the people". Similarly, Nietzsche unmasks religion to reveal it as the refuge of the weak. Likewise with Freud, the same pattern of "unmasking" to reveal and distinguish "the real" from the "apparent" is

evident in his analysis of religion. So, while religion was perceived to be a legitimate source of comfort and hope when one is faced with the difficulties of life, in reality religion was an illusion that merely expressed one's wish for a father-God.

In this respect, my understanding of Ricoeur's dictum is slightly unfaithful to his own. For a false consciousness is no actual experience. Ricoeur himself insisted that it would be a mistake to view the three as masters of scepticism. They are involved with destroying established ideas, not with criticising authentic experience. Quoting Ricoeur himself:

> All three clear the horizon for a more authentic word, for a new reign of Truth, not only by means of a 'destructive' critique, but by the invention of an art of interpreting. (Ricoeur 1970, p. 33)

All three, for Ricoeur, «represent three convergent procedures of demystification» (Ricoeur 1970, p. 34).

Once a false consciousness is demystified, authentic experience can take place again, and reality revealed, within the limits of an age's conceptual and cultural means. In this respect, the masters of suspicions are no masters of scepticism.

Now, independently of Ricoeur's purpose, I do believe that our age *is* an age of scepticism, thereby interpreting the school of suspicion in a much more radical way, namely as a school of complete lack of confidence in the truthfulness of experience itself.

1. Scepticism and Phenomenology

Philosophy of nature as well as philosophy of culture has proposed many reasons to doubt that things are as they appear, over the last century. The "culture of suspicion" – in my radical interpretation – that is a majority of continental philosophers of the twentieth century, on the one side, and the mainstream naturalism striving toward an image of the world compatible with contemporary science on the other side, suggests that our experience (and our moral experience quite particularly) is a pervasive, systematic illusion. They could be right.

Why has this happened? The story would be too long to tell: we shall limit ourselves to pointing to the two mentioned contemporary forms of scepticism

concerning visible things – or the visible and sensible life-world: which we may term as Post-modern Relativism and Reductive Materialism.

The first one has been the dominant philosophy of culture, whereas the second one has been the dominant natural philosophy of man and his mind. Both represent a form of scepticism relative to the immediately given things of our life-world, including ourselves, human persons.

According to post-modernism no real epistemic credit can be given to immediate cognition or consciousness – no form of intuition, acquaintance, perception, feeling is a mode of veridical experience, the world being as it were wrapped up in language, culture, interpretations.

But according to reductive materialism, phenomena are epi-phenomena, just shadows or dreams caused by a completely different reality. Take for example Daniel Dennett's (1991), "the phenomenological garden": we do not find a description of a real scene like this one, or of a fictional one, similar enough to a human life-world of the Twentieth century on earth, but just a list of qualia, or sense data, in three classes:

1. "Experiences" of the outer world, such as views, sounds, smells, sensations of slippery or rough, of warm and cold, and of our body's position;
2. "Experiences" of the inner world, such as imaged views and sounds, memories, ideas and insights;
3. "Experiences" of emotions and feelings.

All that is purely "subjective", that is belonging to what contemporary philosophers of mind call "phenomenal consciousness", the "hard problem" of consciousness, i.e., phenomenal consciousness.

Actually, questioning the reliability of sensory and sensible experience has been a main trend in the history of modern philosophy, starting indeed from Descartes doubt, going on with Galileo and Locke's expulsion of secondary qualities from the furniture of the real world... Yet the "age of suspicion" induced by modern science on the world of everyday experience was at its beginnings in Descartes' days. Nowadays we can perfectly conceive of a world such as that of *Matrix*, where no experienced object is really as it appears: steaks are nothing but tasty *qualia* and people themselves are nothing but the characters of a (shared) dream, while their true life is lived somewhere else...

In fact, the "phenomenological garden" of Dennett, or the world of *Matrix*, is just a set of beautifully arranged qualia, which would support the universal negation of our Principle of epistemic trust:

(N) All appears in vain

(N) supports a version of (epi)*phenomenalism.* And phenomenalism is surely no phenomenology, but the very opposite way of thinking: a radical form of scepticism about phenomena.

Take any issue in contemporary philosophy of mind: the "hard problem" of consciousness, that is the nature of any form of direct cognition, such as perception, emotion, empathy, self-perception; or the nature of the self; or – most important for meta-ethics and legal philosophy – the issue of free will. All of them can be reduced to the general problem of epistemic trust, that is, of reliability of ordinary experience. This is particularly clear with free will.

There is no doubt that we experience free will as the power to determine ourselves to an action, usually in the presence of alternative possible actions; moreover, such an experience seems to be constitutive of our personal and moral identity. Through the decisions I make I assert my identity, stating who I am and projecting the one I shall be – on the background of what I have been. And this is not only true from a first person point of view. I learn to know other people from their actions, through the emotions, the sentiments that their voluntary actions arouse in me: gratitude, grudge, admiration, disdain – and the corresponding value judgments. All the realm of moral experience supposes that we do in fact enjoy free will.

2. Can We Take Experience Seriously?

The question is whether this kind of experience is valid – even though its fallibility, as any other experience of reality – or whether it is systematically deceptive: whether it *can* be veridical or not, whether it *does* correspond to something beyond the experience itself, in reality. This is the general meaning of most philosophical questions today, and free will is just a privileged issue to focus on it.

Now, moral experience is just a part of value-experience (morally good or bad, and all of the virtues and vices, are, respectively, positive or negative values of voluntary actions, or habits). In order to take moral experience seriously, I first have to take value-experience seriously. Morality presupposes

that there are things of value, negative and positive; that there are things and states of affair which are valuable in some respect (pleasant or unpleasant, beautiful or ugly, precious or cheap, holy or unholy etc.), and even more or less valuable.

Moral goodness, in fact, can be defined as the property of a voluntary action (or behaviour, or habits, or intention) aiming at realizing the higher possible value in the given situation.

More specifically, the human world is full of wrongs, for example of killings, frauds, act of violence etc.; moreover, there are lots of things which seem unfair even when there is nobody acting unjustly (e.g., depending on economy or social relations), there are vulgar attitudes and ugly pictures etc.

Am I justified in taking all this experience seriously? That is, in considering experience, including moral and value experience, either as reliable or as at least correctible, in any case as such, that we can learn from it, use it as evidence for our judgements and inferences, etc.? Has our experience generally a cognitive value? And if perception does, why emotion should not?

Let's consider my indignation at a base act, like cheating a defenceless child. In order to take this experience seriously, I must believe: (1) that the agent acted freely, and that free will is no illusion; (2) that the action is actually base, a moral wrong, hence that there are negative or inferior values that the action realizes instead of positive or higher ones.

Hence in order to take my indignation seriously I must entertain a) an ontological b) an axiological belief.

Am I justified in having this kind of beliefs? The question is: can beliefs of this kind be true and justifiable, even if they were not justified in this particular case?

3. Epistemic Trust and Personhood

The answer is yes, only in case (ET) is true. In fact, phenomenology is born to oppose scepticism concerning the phenomenal world, be it of a post-modern relativist, or of a reductive materialist kind.

Why should we adopt epistemic trust instead of scepticism, or phenomenology instead of phenomenalism?

I'll argue that epistemic trust is a necessary condition for human animals to become persons, that is, reasonable or responsible agents. The point of the

argument is that, if I am right, no human animal can become a subject of acts, or develop a selfhood, without entertaining a relation to reality which is a relation of epistemic adequacy – as opposed to simple biological adaptation. In other words, one does not become a normal, autonomous individual of the human kind without entertaining a relation with truth and falsity: a relation which is fundamental even before being voluntary, or conscious.

Let us begin by quoting a passage from a social phenomenologist, Peter L. Berger:

> To become a parent is to take on the role of world-builder and world protector. The role that a parent takes on represents not only the order of this or that society, but order as such, the underlying order of the universe that it makes sense to trust. (Berger 1995, p. 55)

«Everything is in order, everything is all right» (p. 55) – that is the kind of sentence by which any parent reassures her children. This phrase, Berger says, can be expanded into an assertion of cosmic scope: "Be confident. Trust what there is". He goes on:

> This is precisely what the formula intrinsically implies. And if we are to believe the child psychology [...] this is an experience that is absolutely essential to the process of becoming a human person. Put differently, at the very centre of the process of becoming fully human at the core of *humanities*, we find an experience of trust in the order of reality. (Berger 1995, p. 55, 56)

We must be more analytic to understand the deep issue which is at stake in this passage. What is being "built" in the relation between a parent and a newborn child is what phenomenologists call the self-evidence of the life world, or, as Erwin Straus has it, the axiomatic of the everyday world: to sum up, the fundamentals of that shared tacit knowledge, mostly practical knowledge, know how or "sich bekennen", being familiar with, that is *common sense.* Husserl introduces the concept of *transcendental trust*: i.e., the confident expectation that experience keeps going on in the same constitutive style, or according to the same constitutive rules (Formal and transcendental logic). The real world, Husserl underlines, «exists only on the assumption, constantly prescribed, that experience keeps going on in this same constitutive style» (Husserl 1929).[1] L. Binswanger quotes this passage from Husserl in order to emphasize the tragic loss of "natural evidence" *(natürliche Selbstverständlichkeit)* which can take

[1] Quoted by L. Binswanger (1960, p. 24 ; the translation is mine).

place in schizophrenia or major depression, when the patient experiences "the end of the world". Actually this lost of transcendental trust is the loss of "normality" – the loss of reason and even personal identity, the very basis of severe psychopathologies.

John Searle calls "background" this largely shared set of tacit cognitions and abilities which are, according to him, no intentional states (beliefs or intentions), but allow intentional states to refer or to have conditions of satisfaction. This background contains the enormous number of implicit norms, or patterns of "normal" behaviour, that we follow when dressing up (order of suits, socks, shoes) or cutting a cake (one does not cut it like one cuts the grass), even if any explicit direction about how to act correctly is missing. But, as we learn how to behave more or less adequately by "doing with", or taking part in common activities, sharing ordinary life, so we learn how to respond in appropriate ways to events in the environment by sharing experiences, "right" ways of perceiving and feeling.

As flourishing researches in social ontology and social cognition have shown, we – the "neotenic" animals, the ones whose training to autonomous life is the longest one – learn by shared intentionality the right ways to be and act in the world. How do we achieve this apprenticeship of reality?

The key-notion of this account is a concept playing a very basic role in Husserl's phenomenology, namely that of Position *(Stellungnahme)*. What follows can also be read as a commentary of a very deep dictum by Husserl, describing the very nature of personal life: *"Alles Leben ist Stellungnahme"*.

Mental life is usually described as a sequence of mental states. This description, current in contemporary philosophy of mind, is unfaithful to *mental life of a person*. Personal life is no sheer sequence of mental states (such is a dream) but rather a motivational connection of acts. Let me quote two passages by Husserl, where he points out the relation between *positionality* and *normativity* – or, as I would say, "normality" of our mental life:

> Alles Leben ist Stellungnehmen, alles Stellungnehmen steht unter einem Sollen, einer Rechtssprechung über Gültigkeit oder Ungültigkeit, nach prätendierten Normen von absoluter Geltung. Solange diese Normen unangefochten, durch keine Skepsis bedroht und verspottet waren, gab es nur eine Lebensfrage, wie ihnen praktisch am besten zu genügen sei. Wie aber jetzt, wo alle und jede Normen bestritten oder empirisch verfälscht und ihrer idealen Geltung beraubt werden? (Husserl 1987)

In order to understand this passage better, we must recall that central achievement of Husserl's which is his *unified theory of reason* (theoretic, axiological, practical), as the realm of acts subject to normativity, or the distinction right/wrong. Here is a passage nicely summarizing that achievement:

> Der Deutlichkeit halber bemerke ich, dass das Wort Vernunft hier nicht im Sinne eines menschlichen Seelenvermögens, sondern einen Titel für die wesensmässig geschlossene Klasse von Akten und ihre zugehörigen Aktkorrelaten befasst, die unter Ideen der Rechtmässigkeit und Unrechtmässigkeit, korrelativ der Wahrheit und Falschheit, des Bestehens und Nichtbestehens usw. stehen. Soviel Grundarten von Akten wir scheiden können, für welche dies gilt, soviel Grundarten der Vernunft. (Husserl 1988)

This way, the whole set of "*intentionalen Erlebnisse*" – that is *"Akte"*, partitioned into the three classes of cognitive or "doxic", axiologic or "*wertende*", practical or conative "Erlebnisse" are described as subject to normativity. The life of reason starts with the life of a person, permeates all her experiences, perceptions, feelings, intentions, desires, decisions... A very "aristotelian" picture indeed, very far from Cartesian and post-Cartesian dualism of mind and body, reasons and passions etc.

Normativity is an essential feature of intentionality, though a very neglected one both in continental and analytic philosophy of mind: yet it pervades the whole extent of our mental life. This is a deep insight phenomenology offers, suggesting that we should look at personhood as the condition of what we may call "the normative animal". A description of what we mean by "normative animal" can be found in this remarkable passage by Edmund Husserl:

> Das Tier lebt unter bloßen Instinkten, der Mensch auch unter Normen. Durch alle Arten «von» Bewußtseinsakten geht ein damit verflochtenes normatives Bewußtsein von richtig und unrichtig (schicklich, unschicklich, schön, häßlich, zweckmäßig, unzweckmäßig usw.) und motiviert ein entsprechendes erkennendes, wertendes, dinglich und gesellschaftlich wirkendes Handeln. (Husserl 1989)

Consciousness and normativity are essentially bound in our life. Now, how is this possible, from its very beginning? For, according to this description, we do not first perceive, feel or act and only later learn to perceive, feel or act adequately; we are subject to normativity from the very beginning. We experience the world in such a way as to be at least able to learn from our

errors, to correct them. We are bound to be reasonable from the very outset of our life. How is *that* possible?

Husserl's answer to this question sheds light on many peculiarities which distinguish our very early dispositions to social cognition from those of other primates, as described in the pioneering work of Michael Tomasello (1999, 2008 and 2009).

We won't go into details here, but shall only point out to the essential insight Husserl allows us to work out, by linking, as he does, normativity to positionality, this other pervasive and largely neglected feature of intentionality. The upshot of this move is realizing that the exercise of reason is impossible without that of freedom – a pretty radical and yet non-arbitrary kind of freedom, largely unknown in the other animal species on earth. This non-arbitrary kind of freedom is the very basis of personhood, *in the sense that it is constitutive of it.* Hence, there is no exercise of reason without that of personhood. Personhood is no sheer biological condition, neither is it a sheer social status, conferred to us as that of belonging to a community, as being acknowledged as a member in other primates' communities. *Personhood is the more or less adequate exercise of positionality.* It is a biologically grounded disposition which actualizes itself in the progress of adequate position-taking in response to the environment. It is the work of the subjective side of intentionality. Yet this adequacy (right or wrong) cannot be there before we ourselves are there. And "we" are quite apparently not yet there at the very beginning. At the very beginning, our positionality is random, our Stellungnehmen is largely arbitrary. There is a "freedom" which precedes us, so to speak. If this "freedom", or rather arbitrary positionality, is not adequately "guided", we won't develop a "normal" personal life, a life of "reason".

Teaching to take position adequately is the task of the original life-community which welcomes us at our birth, or one fundamental task of parental care – so obvious, that it often goes unnoticed. Only on the basis of a "correct" or truthful relation to factual and axiological reality of the environment can we develop the motivational coherence making up a self or a subject of further experience and action. But what is adequacy or correctness for a baby or a very young child?

Right and wrong – this is the law and ethos of the life community, most originally of the parental care-takers. This is what Berger meant by saying that parents "bring order into the world":

> A child wakes up in the night, perhaps from a bad dream, and finds himself surrounded by darkness, alone, beset by nameless threats. At such a moment the contours of trusted reality are blurred or invisible, and in the terror of incipient chaos the child cries out for his mother. It is hardly an exaggeration to say that, at this moment, the mother has been invoked as a high priestess of a protective order. It is she (and in many cases she alone) who has the power to banish the chaos and to restore the benign shape of the world. (Berger 1970, p. 54)

Mother is right in all she does to assert that there is no danger, that "all is in order". But how can the infant know she is right? Well, this is epistemic trust, the more fundamental and necessary kind of trust. The necessary condition, not only to grow adult, and to verify whether that trust was just or not (maybe nobody of us mortal beings can really verify the absolute truth of that assertion – we only learn to know its relative truth). Epistemic trust is a necessary condition to become a "normal animal", a human person.

4. Some Details

The basis of our entire personal life is given by what we may call *basic acts*, involving *first level positions.*

4.1. First level positionality

There are two classes of such basic acts: cognitive or emotional, perceptions and emotions. Cognitive basic acts, perceptions are characterized by first level "doxic" positionality; emotional basic acts by "axiologic" positionality.

What we call doxic positionality is realizing, taking note of the perceived thing's existence. It is a kind of assent or denial, not a reflexive but an immediate one: yes, the thing is there. A perception *can* turn out to be a delusion. It could *not*, if there were no doxic position, like in an act of imagination or day-dreaming. A doxic position corresponds to the pretense of veridicality which distinguishes perceptions.

What we call axiologic positionality – is realizing the positive or negative salience, or value, of the given thing or situation. Each emotion includes such a position. In fact, emotions can be appropriate, or not. But they could not turn out to be non appropriate – such as panic in front of a very peaceful little cat – if they lacked any axiological position.

First level positions *are not free*. I cannot avoid endorsing the existence of what I see or touch; I cannot take up an opposite position on the negative value of an object of fear, or horror. Even in case the thing turns out to be a delusion as experience goes on, or the fearful beast not to be that bad after all.

What is the role of positionality in basic experience? It should be clear by now. Only positionality is responsible for *adequacy* of perceptions and emotions. Perceptions are veridical or not ; emotions are appropriate or not, *in virtue of their positions.* Hence, if by "experience" we don't mean just causal impact of external reality on an organism, but *something we can learn from*, something which is or is not veridical, something which can provide evidence for our judgments, then we must take positionality into account.

To sum up: (basic) acts are *adequate or inadequate* responses to reality. By adequacy, I mean rational adequacy, in a broad sense: cognitive and practical. Personal life as a life of reason starts with the basic acts. Or, we can also say: basic acts constitute a first level of emergence of a person on her states: the level of evidential objectivation.

4.2. The role of epistemic trust

Now, let's observe a child or a newborn. Consider her basic experiences, emotions and perceptions. In every perception there is something like a yes or a no, an existential proto-judgement. Mother is there – or she isn't. In every emotion there is something like an axiological yes or no. Good and evil, well-being, tummy-ache. Way before being able to voluntary or reflexively position taking, we spontaneously respond to the data of the environment – factual data and/or data of value – that are conveyed by perceptions and emotions. We respond with a sort of cognitive and emotive yes and no.

Initially, though, these positions are largely chaotic: clear in the limiting cases of crying and satisfaction, easy to turn into their opposites, they seem to follow each other as simple states, without a punctual "correspondence" with reality, and without an internal "coherence". The care-giver brings order in the baby's world by reinforcing all (and only) the adequate positions, and the same does the community within which the baby grows up. "Nasty table, it hurts you", says mother while beating the edge of the table, "yeees so good!" – by feeding her child.

Indeed, a child *learns* to take a position – to take a position *correctly*, at the level of basic positions. They are not "free" (for one cannot choose whether or not approving of well-being or crying with tummy-ache) but can be so

inadequate, random and chaotic, that they would prevent the configuration of a unitary subject, with a motivational coherence, memories, and expectations. In order to constitute (truthful) experience, and hence the ground of our life, the pulsation of positional yes and no should not be totally dependent on emotional states, drives and desires. But how do we teach our children right and wrong? We take the right positions *with* them, we share positionality. Only in this way an ordered world, more or less objective and filled with positive and negative qualities, emerges from a flux of sensory, emotional, enactive experience.

We can verify it every day, even observing young humans far beyond the age of what we may call primary self-constitution, or the apprenticeship of the basic skills of personhood, within the customs and language of the concerned life-community. Without a discipline of consents and prohibitions, of positive and negative endorsements on the part of the concerned life-community, no new member of it ever becomes a "normal" subject, a person finally capable of responsibility and reason. A person only grows up on the basis of the right and wrong responses that we learn to give in our infancy – and far beyond. For we humans never stop growing up: "ripeness is all", but it is seldom reached.

A set of socially transmitted rules, a second cultural nature, seems to be a necessary condition for what we called primary self-constitution, the emergence of a "normally" behaving human subject. Epistemic trust is the basic condition for this transmission, and this would conclude the argument.

4.3. Free acts

It would not end the phenomenology of our growing up, though. Personhood involves individual personality. We have so long examined the role of positionality in making up the solid ground of a life capable to learn from experience, indefinitely, and to save acquired knowledge for future generations (as other primates don't do, or very little). Is this its only role?

Of course not, if primary self-constitution is not human or personal ripeness. Personhood is a highly individuated "normal" behaviour. Within the range of normality, there is no function (perception, cognition, memory, emotional life, language) whose exercise would not appear, in our species, highly "personalized". How do individual personalities emerge? Here is a further job for positionality within our intentional life (in the broad sense of "intentional").

Basic positions are not free – it is not in our power to see something that isn't there or to feel as good something that hurts us. But we *can* switch attention from the factual datum, as we can "neutralize" the negative emotional datum, instead of "taking them over" and let us be "motivated" by them to further exploration, further emotions or actions, even in the passive sense of agreeing to an incitement to further experience.

These "removals" and "acceptances" are second order positions. Second order positions generally *are* in our power: they are free acts – in a broad sense of "free", which does not necessarily involve reflective consciousness, let alone deliberation. They are responsible for those spontaneous and largely unreflective (in a sense, "unconscious") strategies of avoidance and pursuit through which everyone track his life in the world, thereby manifesting personal motivational patterns, a "character" or a "personality". Some of us, still in a cradle, pay more attention to colours, other to sounds. This spontaneous and unconscious *management of our passivity*, so to speak, or of our exposure to the experience, manifests a kind of "freedom" – or individualization of behaviour – largely unknown among other primates. The exercise of it is what makes us different from each other. If positionality of the first order, or adequate positionality, constitutes us as reasonable ("normative") animals, positionality of the second order, or free positionality, constitutes us as individual persons. In a sense, this "freedom" precedes and shapes us, as our actions and activities do all over our life.

The object of these second order positions is nothing well defined and structured as a project, not even a meaningful voluntary action like that of comforting a friend or preparing a coffee. They define what we can describe as the *grey zone of spontaneity*. And this grey zone where the human behaviour has a limited responsibility is surprisingly vast. It not only covers early infancy behaviours, it is not only typical for collective behaviours with their sometimes inhuman consequences (the "big animal", said Plato), but it is also the *basso continuo* of our conscious life, the ensemble of its routines, the ground of our "familiarity" with the world and with the others.

It is surprising how much of ourselves, of our individual selves, is "built" in this grey zone of spontaneity, which harbours a part of the enigma characterizing human personality, for better and for worse. Indeed, by exposing and not exposing myself to a certain path of further experience, emotions, actions, I determine "myself", emerging from the states I happen to live in (while other primates just keep living in them) and I orient my life

instead of just living through it. In a certain measure, I make myself responsible for what I become.

4.4. A CONCLUSION ON WILL AND FREE WILL

Positions of this order are in broad sense free acts, but they lack of a *conscious intention* – of a *purpose*. Free will – the conscious exercise of a *power to endorse or not* any given motif of action – desires, drives, aspirations, emotions, interests, engagements, duties – is not yet involved here. Free will, or rather decisions and choices actualizing it – represent a positionality of a further level, by which a possible reason for action is transformed in an actual, causally efficacious one. In fact, what else is "the will" if not *positionality or power of endorsement at this level of cognitive, axiological and practical acts, or "reason"*? It could definitely not exist without the interplay of "normality" and "spontaneity" at the inferior levels, without non-free and free positionality. But once the inferior levels are granted, why should free will not be as real as it seems to be?

Why then does the problem of free will seem so insoluble? Our analysis shows that this depends on a sort of *fallacy in the order of explanation* of the relevant phenomena. Most philosophers presuppose our existence as human persons (without saying in what it is characteristic), and wonder whether our "will" (without explaining what they mean by this word) is "free" (sometimes without really defining this predicate). They don't observe, instead, the two described features of our being:

1. A truthful or at any rate correctible relation to factual and value data of experience, a "normality" of responses
2. A surprising discretionary power through which any human being lets himself get motivated by those data, thereby manifesting what we call her "character".

These features seem to be constitutive conditions of personhood, required for "reason" and "will" to be there too. Only on their basis will decisions and choices become possible, as soon as feasible and meaningful actions can be represented as projects and turned into effective actions by decisions. As self-obligations, decisions and choices are self-constitutive acts at a higher level, in which identity through time is constituted and modified: since any such project involves taking over responsibility for one's future self and recognizing oneself

responsible for past actions. Any decision involves a conscious endorsement or reject of what we are already. A decision involves a first person reflective attitude, something far beyond the spontaneous management of one's passive states. Much more basically than in the exercise of free will, phenomenology opens up the interplay of chance, norms, freedom and truthfulness through which we build ourselves as the persons we shall be, by trial and error.

REFERENCES

Berger, P. L. (1970). A *Rumor of Angels: Modern Society and the Rediscovery of the Supernatural*. New York: Anchor Book.

Binswanger, L. (1960). *Melancholie und Manie – Phaenomenologische Studien*. Pfullinger: Gunther Neske.

Dennett, D. (1991). *Cosciousness Explained*. Boston: Little, Brown and Company.

Husserl, E. (1929). *Formale und transzendentale Logik. Versuch einer Kritik der logischen Vernunft*. In E. Husserl, *Jahrbuch für Philosophie und phänomenologische Forschung*; vol. X. Tübingen: Max Niemeyer Verlag.

Husserl, E. (1987). *Philosophie als Strenge Wissenschaft*. In E. Husserl, *Aufsätze und Vorträge (1911-1921)*, Hua XXV. Dordrecht: Kluwer Academic Publisher.

Husserl, E. (1988). *Vorlesungen über Ethik und Wertlehre 1908-1914*, Hua XXVIII. Dordrecht: Kluwer Academic Publisher.

Husserl, E. (1989). *Fünf Aufsätze über Erneuerung*; *Formale Typen der Kultur in der Menschheitsentwicklung*. In E. Husserl, *Aufsätze und Vorträge (1922-1937)*, Hua XXVII. Dordrecht: Kluwer Academic Publisher.

Ricoeur, P. (1970). *Freud and Philosophy: An Essay on Interpretation*. New Haven: Yale University Press.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M. (2009). *Why We Cooperate*. Cambridge, MA: MIT Press.

Perspectives on the Experience of Will

Davide Rigoni *
davide.rigoni@unipd.it

Luca Sammicheli **
luca.sammicheli@unibo.it

Marcel Brass ***
marcel.brass@ugent.be

ABSTRACT

In the last decades, psychologists and neuroscientists brought the concept of human will out of the philosophical debate. Here we critically examine the different attempts within the field of cognitive neuroscience to study neural processes underpinning human will. Volition has been investigated under different perspectives: while some threads of research focused on the subjective experience of free will (i.e., will under a self perspective), others explored how the brain is able to identify free will in other individuals (i.e., will under a other perspective). In addition, we comment that perceiving free will in others is tightly connected to the ethical and juridical concept of personal responsibility. Finally, we present a promising theoretical framework that stresses the pragmatic value of believing in free will. Rather than focusing on the subjective experience of volition itself, this approach studies whether believing in free will or not has an impact on brain processes underlying willed behaviour.

1. Introduction

The subjective feeling of controlling our own actions is an intuitive and pervasive component of human experience. When switching on the TV to

* Department of Developmental Psychology and Socialization – University of Padua
** Faculty of Psychology – University of Bologna
*** Department of Experimental Psychology – University of Ghent

watch the news or when entering a pub to order a cappuccino, we have the clear feeling of voluntarily and freely determining our choice. The question of how we can voluntarily control our behavior has always fascinated researchers from different disciplines such as philosophy and psychology. This question is fundamental to what it means to be a human being and is tightly related to socially relevant issues, such as personal responsibility and self-control.

The fascination for willed behavior is to some degree fuelled by the *vexata quaestio* of free will. In the last decades, cognitive neuroscientists and experimental psychologists focused on intentional actions, sometimes assuming – more or less explicitly – that understanding brain processes involved in conscious and voluntary actions (i.e., those actions that we perceived as *free*) would provide an answer to the question whether free will exists or not, or at least would modify our notion of volition. However, it is highly questionable whether the fields of neuroscience and experimental psychology have tools for answering the question whether free will, in philosophical terms, exists.[1] As Roskies concluded in her recent review, «neuroscience has not much affected our conception of volition [...]» but «[...] it has typically challenged traditional views of the relationship between consciousness and action» (Roskies 2010, p. 123).

Therefore, the present paper will focus on the neural mechanisms underlying the *subjective experience of free will or volition* without trying to relate these findings to the philosophical problem of free will.

In the first part, we will critically discuss a series of empirical findings within the field of cognitive neuroscience that explored what brain mechanisms precede the experience of free will. These findings have strongly influenced the notion of the relationship between consciousness and intentional actions. In the second part, we will examine the *reconstructive* approach of the experience of will. According to this perspective, our experience of volition is strongly influenced by events occurring after the action is executed and sometimes is retrospectively reconstructed. Then we will briefly discuss how we perceive free will in others. This part of the paper will outline the processes that underpin our ability to identify intentionality in other individuals. In addition, we will describe how tightly these processes are related to ethical and juridical issues. Finally, we conclude by presenting a recent theoretical framework that stresses the pragmatic value of believing in free will. Rather than focusing on

[1] For a recent review, see Roskies 2010.

the subjective experience of volition itself, this approach studies whether it has any implication whether we believe in free will or not.

2. The Subjective Experience of Volition: Cause or Consequence

From a phenomenological point of view, we may define as *free* those actions that are performed intentionally and with a minimum of external constrictions. When we have the intention to perform a specific action, we feel that our intention is, somehow, *causing* the action itself; in other words we feel that our action is *determined* by our intention to perform that action. We refer to this feeling of willing as *conscious intention* (Haggard 2005).

A first line of research within the field of cognitive neuroscience has focused on whether the subjective experience of free will plays a causal role in the initiation of behaviour. In a pioneering experiment, Benjamin Libet and colleagues (Libet *et al.* 1983) applied neurophysiological methods to study the relationship between the electrophysiological brain activity associated with voluntary movements and conscious intentions. The main interest was on the *temporal relationship* between motor-related brain potentials, as recorded with the electroencephalogram (EEG), and the 'conscious feeling of intending to act'. Thus, the question was: *when* do people become aware of their own decision to do a certain movement? And what happens in the brain in the meantime?

An implicit problem in investigating internal representations such as the conscious intention to perform a movement, is that it is impossible – at present, at least – to obtain a direct and objective measure of *when* a person becomes aware of his or her conscious intention. It is not possible to have a direct access to the 'internal world' of others and therefore, to obtain an estimation of when people had the conscious intention to execute a movement, experimenters must rely on introspection (i.e., subjective reports of inner states). Libet and colleagues (Libet *et al.* 1983) developed a method that allowed to compare subjective self-reports with brain activity. In the experiment, participants were seated in front of a screen displaying a clock with a rapidly moving spot and they were asked to execute a rapid movement (i.e., a wrist flexion), at will. Afterwards, they were asked to report what time it was (i.e., the position of the spot in the clock) when they had the first subjective experience of intending to act (see *Figure 1*). Libet referred to this reported

time as the will judgment (W). At the same time, movement-related cortical potentials were recorded by means of a surface electrode placed on participants' scalp.

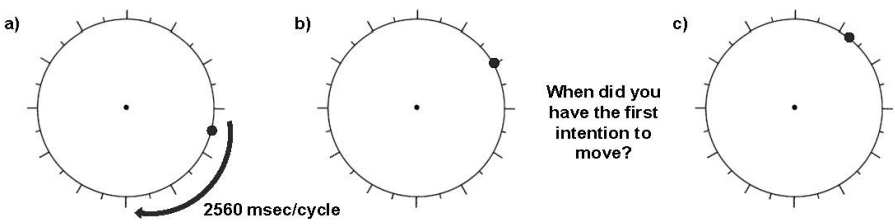


Figure 1: A typical Libet's clock paradigm is represented. (a) Participants make a voluntary and spontaneous finger movement while watching a cursor moving clockwise around a clock face. (b) At a variable time after finger movement, the cursor stops. (c) Then, participants are asked to report the position it was when they had the first intention to make the movement.

Libet was mainly interested in a well known cortical potential, the *Bereitschaftspotential* or *readiness potential* (RP) (Kornhuber and Deecke 1965). The RP is a slowly increasing negative potential which starts up to 2 seconds before voluntary and spontaneous movements and is bilaterally symmetrical over the pre- and post-central region, with a maximum at the vertex (Shibasaki *et al.* 1980, Shibasaki and Hallett 2006). The RP is generated by the supplementary motor area (SMA) – including pre-SMA and SMA proper –, a brain region involved in the late stages of motor preparation (Shibasaki and Hallett 2006). It is associated with spontaneous and voluntary movements and it is absent or greatly reduced before involuntary movements or movements made in an automatic manner (Shibasaki and Hallett 2006).

In the original experiment (Libet *et al.*, 1983), participants' voluntary movements were preceded by a RP beginning 500 ms to about 1000 ms before movement onset. The W-judgment, indicating the time when people had their first intention to move, was approximately 200 ms before the motor response. Therefore, the brain potentials reflecting motor preparation began about 300 to 800 ms *before* the person consciously intended to act. Conscious intentions would thus seem, the authors concluded, to be a latecomer in the process of

decision, rather than the generator of the action.

Several theoretical and methodological aspects of the Libet's clock paradigm have been extensively questioned (Hallett 2007, Pockett and Miller 2007, Roskies 2010). However, despite the numerous theoretical and methodological critiques[2], the Libet's clock has been widely used to investigate conscious intentions and it still offers «one of the few viable methods for experimental studies of awareness of action» (Haggard 2005, p. 291). Furthermore, the main result of Libet's experiment has been repeatedly confirmed by other empirical studies that clarified the temporal relationship between conscious intentions and brain processes underlying motor preparation. For instance, Haggard and Eimer (1999), replicated the original findings and found that the reported W correlates only with the late part of the RP – i.e., the lateralized RP – that represents the stage at which the representation of an abstract action is translated into representation of a specific movement (i.e., "Do that!"). This finding suggests that we become aware of our own intention to perform a voluntary movement only when information about which specific movement has to be made is represented in pre-motor areas (Haggard and Eimer 1999).

These data show that our motor actions are preceded by preconscious brain activity, which enters our awareness only at a later stage, just before the action is executed. Therefore, a plausible conclusion is that conscious intentions are not the first source of our behavior as voluntary actions would be primarily determined by brain activity that enters consciousness only at the later stages (Hallett 2007).

However, these conclusions are far from being uncontroversial. For instance, Trevena and Miller recently questioned the assumption that the RP is specifically associated with voluntary movements (Trevena and Miller 2010). They thought to show that the RP is not necessarily followed by an overt movement and therefore it cannot be considered a *specific* marker of voluntary movement preparation. However, their experimental setup has also been criticized (Gomes 2010). Therefore, further research is needed to better clarify the relationship between brain processes underlying voluntary movements preparation and the subjective experience of intention.

It is noteworthy to mention that more recent studies extended the hypothesis that our behavior is determined by unconscious brain activity using

[2] See Haggard 2008; Hallett 2007; Pockett and Miller 2007.

functional MRI. For instance, Soon and colleagues (Soon *et al.* 2008) used a modified version of the Libet's clock in which participants had to freely decide between a left and a right button press that they should execute at a freely chosen time. Then, participants reported the time at which the conscious motor decision was actually made. They found that the outcome of the decision – i.e., whether the left or the right button – was encoded in the brain activity of prefrontal and parietal cortex up to 10 seconds before it entered awareness. The decoding accuracy was about 60%. Thus, these data show that brain activity preceding awareness can predict our conscious decisions[3]. It is crucial to note here, however, that the prediction even though reliable was far from being perfect. With an accuracy of 60%, that is, 10% above chance, it is difficult to argue that these information *determine* our decision. This finding does not tell us that our conscious decisions are fully determined by such unconscious processes; rather it indicates that our conscious decisions are *biased* by brain activity reflecting unconscious processes. One crucial question is whether the low accuracy is due to methodological shortcomings or to principle reasons, namely that the bias is simply not stronger than, for example 10 %. It would be interesting to further investigate whether it is possible to influence the accuracy of the prediction.

3. Reconstruction of Intentions and Apparent Mental Causation

In voluntary actions we experience that the conscious intention to perform an action precedes the action itself. Subjectively, the intention to press a key *determines* or *causes* the key press. A series of empirical studies in experimental psychology and neuroscience attempted to challenge this intuitive experience by focusing on cognitive and brain mechanisms underlying the evaluation of the consequences of our actions, as these processes seem to influence the subjective experience of conscious intentions.[4] Empirical data suggest that the subjective experience of the conscious intention is strongly influenced by events occurring *after* the action

[3] Indeed, what is *unconscious* is not the brain activity itself, but the mental state associated with that brain activity.

[4] See Banks and Isham 2009; Kühn and Brass 2009; Lau *et al.* 2007; Rigoni *et al.* 2010; Wegner and Wheatley 1999.

is executed. Conscious intentions would then be, at least partially, retrospectively *inferred* from events occurring after an action is executed.

A study by Lau and colleagues (Lau *et al.* 2007) provided evidence in favour of this *reconstruction hypothesis*. They applied a Transcranial Magnetic Stimulation (TMS) over the pre-supplementary motor area (pre-SMA) after the execution of a simple spontaneous movement while participants were performing a Libet's task. They found that when the TMS pulse was applied 200 ms after movement execution, the perceived onset of the conscious intention shifted backward in time, indicating that the experience of conscious intentions involves activity of the pre-SMA taking place after the execution of action.

Banks and Isham (2009) used a modified version of the Libet's procedure in which participants were asked to press a button at will and to report the W judgment – i.e., the time they had the intention to press the button. Immediately after each button press, an auditory feedback was delivered at variable delays of 5, 20, 40, or 60 ms, in order to signal a response later than the actual one. Although participants were not aware of the delay, their W judgment moved forward in time linearly with the delay of the auditory feedback, indicating that people estimate the timing of their conscious intentions on the basis of the apparent time of response, rather than the actual response. In other words, people estimate the timing of their conscious intentions on the basis of the consequences of the actions, rather than the intention itself.

Rigoni and colleagues (Rigoni *et al.* 2010) extended these findings by applying electrophysiological recordings to the procedure used by Banks and Isham (2009) in order to investigate the psychophysiological mechanisms involved in the inferential processes of the conscious intentions. The authors demonstrated that the inferential processes by which the intention is reconstructed involve brain processes related to action-monitoring.

Taken together, these empirical findings show that the effects of intentional actions have an impact on the subjective experience of free will – at least on the subjective estimation of *when* participants had the intention to act. In addition, they challenge the intuitive view that voluntary actions are caused by the conscious intention to perform that specific action.

Other studies moved a step further and provided evidence that people may retrospectively reconstruct the experience of volition for actions that are executed unintentionally. For instance, Kühn and Brass (2009) combined a

stop-signal paradigm and an intentional action paradigm: participants were asked to press a button as fast as possible when a stimulus, say a letter, was displayed on a computer screen (*primary response* trials). Sometimes, right after the stimulus, either a stop-signal or a decision-signal was presented: with the stop-signal, participants had to inhibit the pending response, with the decision-signal they could decide whether responding to the stimulus or aborting the pending response (*decide* trials). In the decision trials in which participants provided a response, participants were also asked whether it was a voluntary response or a failed inhibition – i.e., participants were not able to stop the response. The aim of the study was to compare the reaction times (RTs) in the *decide* trials in which the subjects decided voluntarily to press the button with RTs in *primary response* trials in order to explore whether subjects were able to discriminate between acting without being able to stop (i.e., failed inhibition) and deciding voluntarily to resume the prepared action. If participants were able of distinguishing those states, there should be no *decide* trials in which subjects stated to have chosen voluntarily to resume the prepared action in the range of *primary response* RTs. That was because the process of stopping an ongoing action and reinitiating it voluntarily should take time. On the basis of this RT analysis, the authors showed that participants judged as voluntary responses that were in the time range of primary response RTs and were thus given unintentionally (i.e., failed inhibitions). Therefore, in some cases, participants had the experience of a conscious decision for unintentional responses.

A more radical view, proposes the so-called theory of *apparent mental causation* (Wegner and Wheatley 1999). According to this hypothesis, people feel that their conscious intentions are the source of their actions because they think about that action in advance of its occurrence, and because alternative sources of the action are not available. The human mind would assume a causal path from the intention to act to the action itself in order to explain the correlation between them (Haggard 2008). This correlation occurs because both the subjective experience of intention and the action are generated by a common process, that is the neural preparation of the movement. Several studies support the idea that sometimes conscious will is fabricated from the *perception* of a causal link between the thought and the action. For instance, Wegner and Weathley (1999) demonstrated empirically that people have the subjective experience that they performed intentional actions that were actually performed by another person. As Wegner commented, «conscious will is not

inherent in action» (Wegner and Weathley 1999, p. 11): conscious intention is not an intrinsic part of the process by which somebody acts, but it is an extrinsic accompaniments to that process.

Taken together, all these studies provide evidence that the experience of volition is biased by factors concerning the consequences of our behaviour. According to some authors, volition is a perception, rather than the generator of behavior. According to this model of free will, our brain motor's system would produce a movement as a product of its different inputs and would inform consciousness of the movement, that would be perceived as being freely chosen (Hallett 2007).

However, one has to be careful with drawing to far reaching conclusions from studies showing that our experience of will is sometimes illusionary. Arguing that free will is *always* an illusion on the basis of experimental observations that it is possible to generate an illusionary *will*, is, in our opinion, an overstatement. Indeed, it is like claiming that our visual system is delusional on the basis of demonstrations of visual illusions such as the Kanizsa triangle or the Müller-Lyer illusion.

4. Experiencing Free Will in Others

Imagine yourself sitting in a crowded bus. Suddenly the bus driver hits the brakes and the bus comes to an immediate stop. The person standing in front of you looses balance and falls on top of you. You feel pain and you are quite annoyed. However, despite a first impulse to react, you feel that a much more appropriate response is to say: "Don't worry, it happens!".

As indicated by the example above, we do not only feel that we are free; we also have a clear feeling that other people are free to act. In other words, as we have an *immediate* subjective experience of free will, we also have an *immediate* subjective experience of others' free will (Gallagher and Zahavi 2008). This ability to immediately and effortlessly discriminate between actions performed intentionally and actions performed unintentionally has been referred to as *intentional stance* (Dennett 1987).

The subjective experience of other people's free will is so instinctive and pervasive that virtually all human societies have formalized it into the juridical category of *personal responsibility*. Personal responsibility is an almost universal concept that is grounded on the ability to identify others' intentions:

the question of 'guilty' vs. 'innocent' actions is meaningful only if we consider the possibility to distinguish between *free* or *intentional* actions and *unintentional* actions.

Among the few studies that focused on the psychological mechanisms supporting juridical categories – in a perspective that in philosophy of law may be called *jus naturalism* – Hamilton tried to describe the *parallelism* between the juridical categories of personal responsibility and the Heider's levels of causal attribution (Heider 1958). According to Hamilton (1978), legal responsibility rules are approximately analogs to the Heider's responsibility attribution levels. For instance, the *association* attribution, in which a person is «held responsible for each effect that is in any way connected with him or that seems in any way to belong to him» (Heider 1958, p. 113), is equivalent to the Vicarious responsibility rule (e.g., regulations that tavern owners are responsible if liquor is served to minors, with or without the owner's knowledge or consent). Similarly, *intention* attribution – i.e., «only what a person intended is perceived as having its source in him» (Heider 1958, p. 113) – is typical criminal responsibility for an intended act (Hamilton 1978).

In law, the use of the different categories of personal responsibility requires the decoding of social behaviour (e.g., a crime) through *mind* constructs (e.g., the intention). That is, the implicit principle of personal responsibility is made explicit by the law in order to distinguish between a *signifier* (e.g., a punch) from a *non-signifier* action (e.g., an automatic reflex in the Tourette's syndrome).

What are the mechanisms by which our brain can distinguish *free* from *determined* actions? Whereas Libet focused on the problem of free will under a *self* perspective (i.e., the experience that '*I* have free will), here the problem is framed under an *others* perspective (i.e., the experience that '*others*' have free will). As outlined in the previous paragraph, attribution of intentionality is crucial for social interactions and for the regulation of human societies, as demonstrated by the existence of the categories of personal responsibility in the law. The study of social cognition – i.e., the processing of information related to the other human beings – is the mean by which the problem of free will – in the *others* perspective – can be investigated. The question moves from the description of the factors influencing the experience of free will to the investigation of cognitive and the neural processes underlying the attribution of free will to others.

Within the field of cognitive neuroscience, different hypotheses have emerged to describe brain mechanisms underlying our ability to attribute free will to others. However, all the different hypotheses rest on the assumption that the first step in the attribution of intention is the ability to distinguish biological from non-biological agents. That is, people must first classify interactions between objects as mechanical or intentional and discern the resence of *agents*, starting from perceptual information (Frith 1999). The brain network underlying the processing of biological motion involves the superior temporal sulcus and the premotor cortex (Beauchamp *et al.* 2002, Grossman and Blake 2002).

The ability to detect agency from biological motion (i.e., psychological causation or intentional movement) is considered a precursor of intentionality attribution. When we observe a biological motion, we attribute mental states to the observed movement, such as goals, intentions, desires. However, we would not attribute intentions to all biological agents but limit it, with a few exceptions, to human agents. Thus, perceiving free will in others requires the ability to understand also other people's goals and intentions. There are two competing hypotheses explaining how we are able to attribute intentionality to others (Gallese and Goldman 1998). The *simulation theory* suggests that people use their own mental mechanisms to predict the mental processes of others. According to the simulation theory, people simulate others' cognitive processes by deploying the same cognitive mechanisms. Conversely, the *theory* suggests that people understand others' intentions by acquiring a commonsense theory of mind, something similar to a *scientific* theory. In other words, people use inferential and deductive processes that do not involve simulation. The two processes involve distinct brain circuits: *simulating* involve premotor and parietal areas, the insula, and the secondary somatosensory cortex, while *theorizing* involve midline structures and the temporal-parietal junction (Keysers and Gazzola 2007).

It has been proposed that the two views describe different types of social interactions that are at the two extremes of a intuitive/reflective continuum (Keysers and Gazzola 2007, Uddin *et al.* 2007): simulationists focus on more intuitive examples in which intentionality is easily and effortlessly identifiable (e.g., when we observe a hand grasping a mug); investigators of the *theory theory* would be concerned with more reflective examples of intention attribution, in which the attribution of intention follows a conscious browsing

through what we know about the observed person and the context (e.g., when someone steps on our toes in a crowded bus) (Brass *et al.* 2007).

The discovery of the *mirror neurons* (Rizzolatti *et al.* 1996) provided an important insight into the brain mechanisms that might be involved in the attribution of others' intentionality. Mirror neurons are a special class of neurons in premotor areas that fire when we perform object-directed actions such as grasping, tearing, manipulating, holding, but also when we observe somebody else performing the same class of actions. Recent empirical findings indicate that the mirror neuron system may be involved also in goal and intention understanding (Hamilton and Grafton 2006, Iacoboni *et al.* 2005), but the involvement of the mirror system might be limited to intuitive situations, as outlined above.

An interesting approach is to link the mirror neuron system with the concept of *semantic nature of human behaviour* (Hauser 2006 and Rawls 1971), in which the *freeness* of a certain action *is* a semantic attribution that leads to an immediate and unavoidable perception of intentionality – "You are free!". Gallagher and Zahavi (2008) propose a theory of social cognition that emphasize the *immediacy* of the attribution of intentionality. This perspective is distinct from the two other main theories of social cognition – the *simulation theory* and the *theory theory*. According to the authors,

> Mirror activation, on this interpretation, is not the initiation of simulation; it's part of a direct intersubjective perception of what the other is doing. At the phenomenological level, when I see the other's action or gesture, I see (I directly perceive) the meaning in the action or gesture. (Gallagher and Zahavi 2008, p. 179)

This approach seems to be well supported from empirical findings on the mirror neurons in social contexts.[5]

Further research within neuroscience is needed to clarify how our brain *perceive* free will in others. For instance, Liepelt and colleagues (Liepelt *et al.* 2008) found that reasoning about the action and the context in which the action is performed have a strong impact on the brain processes underlying the attribution of intentionality to others. This suggests that the attribution of free will to others might be a prerequisite for the activity of the mirror-neuron system, rather than its consequence (Liepelt *et al.* 2008).

[5] See Gallagher and Zahavi 2008 for a review.

However, the attribution of mental states – such as intentionality – to others include other mechanisms as well, namely mechanisms that allow to distinguish one's own intentions from others' intentions.[6] This mechanism involves the right inferior parietal cortex in conjunction with prefrontal cortex.

5. The Pragmatic Value of Believing in Free Will

A totally different perspective on free will comes from social psychology in which human will is viewed as a kind of organ that is fuelled by willpower (Baumeister 2008). This perspective defines human will as a unitary concept that is characterized by specific properties. One central assumption of the *willpower metaphor* is that it draws on a common limited resource (Baumeister *et al.* 1998; Vohs and Schooler 2008). Tasks that require willpower include self-control, decision making, complex problem solving and conflict resolution. From this perspective there is not one task that measures the free will but rather a number of tasks that draw more or less on this resource. In a series of studies, Baumeister and colleagues could show that different tasks requiring willpower indeed interfere with each other (e.g., Baumeister *et al.* 1998; Muraven and Baumeister 2000). More specifically, they could show that carrying out a task that strongly relies on willpower leads to a depletion of this resource – this process is called *ego-depletion* and results in impaired performance in other tasks that rely on willpower. For instance, carrying out a self-control task leads to less persistence in a difficult problem solving task. Furthermore, making free choices to perform attitude relevant behavior also leads to reduced persistence in the problem solving task.

A second basic assumption of the willpower metaphor is that willed behavior is very effortful and requires more energy than behaviour that does not rely on willpower (Gailliot and Baumeister 2007; Gailliot *et al.* 2007). Support for the idea of higher energy requirements for processes involving willpower stems from the observation that such processes are very sensitive to the glucose level (Gailliot and Baumeister 2007).

Given that willed behaviour is so demanding, why do people put so much effort into their behaviour? Why do they spend so much energy to control themselves? Why do they behave responsibly instead of letting their automatic

[6] See Decety and Sommerville 2003 for a review.

and selfish impulses drive their actions? It has been demonstrated that increasing people's sense of responsibility can shift their behaviour toward a more desirable performance (Harmon-Jones and Mills 1999, Mueller and Dweck 1998). Under this perspective, one might expect that reducing people's sense of responsibility may promote undesirable behavior. What would happen if people start to believe that they have no control over their own actions? In other words, what would happen if people would be induced to believe the subjective experience of free will is completely illusional? To address this question, Vohs and Schooler (2008) carried out a study in which they examined whether inducing participants to believe that human behavior is predetermined would encourage cheating. Two groups of participants were exposed either to a deterministic (i.e., statements claiming that high-minded people now agree in that free will is an illusion) or to a neutral message (i.e., statements about consciousness which did not discuss free will). Afterwards, participants were given a series of mental-arithmetic problems. They were told that due to a computer glitch, the correct answer would appear on the screen while they were attempting to solve the problem and that they could stop the answer from being displayed by pressing the space bar after the arithmetical problem appeared. Furthermore, they were told that although the experimenter would not know whether they pressed the space bar, they should try to solve the problem honestly. Unbeknownst to the participants, the dependent measure was indeed the number of times they pressed the space bar to prevent the answer from appearing. Results showed that the participants who were exposed to a determinist message cheated more frequently than those who were exposed to a neutral message. In the same study, the authors showed that also when the task requires a more active behavior in order to cheat (i.e., stealing money from the researchers), participants exposed to a deterministic message behave more immorally than others.

Baumeister and colleagues (Baumeister *et al.* 2009) extended these findings into a broader context. More precisely, they showed that a disbelief in free will increases antisocial attitudes such as aggression and at the same time reduces pro-social behavior such as helpfulness.

These studies show that inducing a deterministic perspective that denies free will strongly influences human behaviour in social contexts. A simple exposure to a deterministic worldview increases the probability that people behave immorally and antisocially. What are the mechanisms underlying this

antisocial shift? Why do people behave antisocial if they are induced to believe that they are not free? According to Baumeister,

> Feelings of responsibility and accountability may make people feel that they ought to behave in socially desirable ways, such as performing prosocial acts of helping and restraining antisocial impulses to aggress against others. The deterministic belief essentially says that the person could not act otherwise, which resembles a standard form of excuse ("I couldn't help it") and thus might encourage people to act in short-sighted, impulsive, selfish ways. (Baumeister *et al.* 2009, p. 261)

Therefore a deterministic message acts as an implicit cue that let people behave in a selfish, impulsive, less altruistic, and aggressive fashion.

One alternative perspective of how beliefs about free will might affect social behaviour is to assume that disbelief in free will changes basic motor cognitive processes which in turn influence how we experience the consequences of our behaviour. Recently, the research group of Marcel Brass attempted to investigate the impact of disbelieving in free will on the preparation of intentional motor action. In particular, they applied the free will manipulation to study brain processes related to the preparation of voluntary movements. They could show that brain potentials that precede voluntary movements and that reflect the intentional involvement in action preparation, are strongly modulated by the level of disbelief in free will (Rigoni *et al.*, submitted). A potential explanation for this result is that the free will manipulation affects intentional involvement in the task via a reduction of self-efficacy beliefs. Less intentional involvement in an action might on the other hand reduce the feeling of agency for the consequences of the behaviour which in turn might alter our experience of responsibility for such actions. Although the specific mechanisms underlying this effect are not clear, these results suggest that abstract belief systems might have a an impact on very fundamental brain processes.

Whereas the studies in social psychology and cognitive neuroscience are crucial in showing the benefits of believing in free will at a societal level another question is how disbelieving in free will can influence individual well being. Promoting the idea that one has few control over his or her own behaviour has a strong impact on how individuals perceive themselves, for instance by lowering individual well-being and by increasing feelings of powerlessness and dissatisfaction.

Believing that we have free will or in other words that we have control over our own actions and over the environment thus seems to be a psychological and biological necessity.[7]

6. Conclusive Thoughts

The subjective feeling of free will is a pervasive component of human experience. We have a clear and unavoidable experience of voluntarily control a great part of our actions and we feel to be the *agent* of our behaviour. We therefore feel we are responsible for those actions that are performed with a conscious intention, that is, those actions that are associated with the subjective experience that "I" decided to do so. However, the *neuroscience of will* (Haggard 2008) has challenged this intuitive experience by questioning the role of free will as the generator of our actions. Here, we critically analyze the most important contributions that have threatened the existence of free will from a neuroscientific perspective. We commented that these studies will hardly provide an answer to the philosophical question of whether free will exists.

Furthermore, we outlined two additional perspectives on free will, namely, how people perceive free will in others and the pragmatic value of believing in free will. Both these frameworks are of great social relevance: human societies are ruled on the concept of *personal responsibility* and therefore it is assumed that people can freely decide their own actions. Understanding the mechanisms underlying the ability to perceive others' intentionality and how disbelief in free will alters basic brain processes, would shed light on several essential aspects of all human societies.

References

Banks, W. P., & Isham, E. A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science, 20*(1), 17-21.

[7] See Leotti *et al.* 2010 for a review.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*(5), 1252-1265.

Baumeister, R. F. (2008). Free will in scientific psychology. *Psychological Science, 3*(1), 14-19.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron, 34*(1), 149-159.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology, 17*(24), 2117-2121.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*(12), 527-533.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Frith, C. D. (1999). Interacting minds-A biological basis. *Science, 286*(5445), 1692-1695.

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review, 11*(4), 303-327.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., & Brewer, E. L. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology, 92*(2), 325-336.

Gallagher, S., & Zahavi, D. (2008). *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. London: Routledge.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the mind-reading. *Trends in Cognitive Sciences, 2*(12), 493-501.

Gomes, G. (2010). Preparing to move and deciding not to move. *Consciousness and Cognition, 19*(1), 457-459.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35*(6), 1167-1175.

Haggard, P, & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*(1), 128-133.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences, 9*(6), 290-295.

Haggard, P. (2008). Human volition: Towards a neuroscience of will. *Nature Reviews Neuroscience, 9*(12), 934-946.

Hallett, M. (2007). Volitional control of movement: The physiology of free will. *Clinical Neurophysiology, 118*(6), 1179-1192.

Hamilton, A. F. de C., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *The Journal of Neuroscience, 26*(4), 1133-1137.

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology, 41*(4), 316-328.

Harmon-Jones, E., & Mills, J. (1999). *Cognitive Dissonance: Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association.

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*: New York: Ecco Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences, 11*(5), 194-196.

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflügers Archiv European Journal of Physiology, 284*(1), 1-17.

Kühn, S., & Brass, M. (2009). Retrospective construction of the judgement of free choice. *Consciousness and Cognition, 18*(1), 12-21.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience, 19*(1), 81-90.

Leotti, L. A, Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences, 14*(10), 457-463.

Libet, B., Gleason, C. A, Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain, 106*, 623-642.

Liepelt, R., Cramon, D. Y. V., & Brass, M. (2008). What is matched in direct matching? Intention attribution modulates motor priming. *Journal of Experimental Psychology. Human Perception and Performance, 34*(3), 578-591.

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33-52.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126*(2), 247-259.

Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition, 16*(2), 241-254.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rigoni, D., Brass, M., & Sartori, G. (2010). Post-action determinants of the reported time of conscious intentions. *Frontiers in Human Neuroscience, 4*, 38.

Rigoni, D., Kuhn, S., Sartori, G., & Brass, M. (submitted). Inducing disbelief in free will alters brain correlates of preconscious motor preparation.

Rizzolatti, G., Fadiga, L., Gallese, V, & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research. Cognitive Brain Research, 3*(2), 131-141.

Roskies, A. L. (2010). How does neuroscience affect our conception of volition? *Annual Reviews of Neuroscience, 33*, 109-130.

Shibasaki, H, Barrett, G., Halliday, E., & Halliday, A. M. (1980). Components of the movement-related cortical potential and their scalp topography. *Electroencephalography and Clinical Neurophysiology, 49*(3-4), 213-226.

Shibasaki, H., & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology, 117*(11), 2341-2356.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience, 11*(5), 543-545.

Trevena, J., & Miller, J. (2010). Brain preparation before a voluntary action: Evidence against unconscious movement initiation. *Consciousness and Cognition, 19*(1), 447-456.

Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences, 11*(4), 153-157.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49-54.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation. Sources of the experience of will. *The American Psychologist, 54*(7), 480-492.

Initiation of Intentional Actions and the Electromagnetic Field Theory of Consciousness

Susan Pockett *
s.pockett@auckland.ac.nz

ABSTRACT

The electromagnetic (EM) field theory of consciousness proposes that consciousness is identical with certain brain-generated EM field patterns. It initially seemed to be a point in favour of this theory that EM fields are known to affect neurons, which in principle provides a mechanism whereby consciousness could affect its brain. However, it is shown here that the sorts of EM fields proposed by the theory as being conscious can act only on neurons that are either identical with, or spatially close to, the neurons that generated those fields in the first place. This makes it difficult to see how putatively conscious EM fields could initiate bodily movements. It does not harm the EM field theory of consciousness however, because an accumulation of independent psychological and physiological evidence shows that consciousness itself is not the proximal cause of voluntary movements. The fact that humans are not directly conscious of the initiation of their own bodily movements is now used to reveal a basic structural feature that may distinguish conscious EM fields from the superficially similar fields produced by those parts of the brain that do not generate conscious experiences.

The electromagnetic (EM) field theory of consciousness (Pockett 1999, 2000, 2002; McFadden 2002a, 2000b) is an identity theory which proposes that consciousness is identical not with particular spatial patterns of neuronal activity *per se*, but with the extracellular EM field patterns that are induced by those spatial patterns of neuronal activity. The major difference between the EM field theory of consciousness and what might now be called the classical

* Departments of Physics and Psychology – University of Auckland (New Zealand)

neural identity theory of Place (1956), Feigel (1958) and Smart (1959) is that the EM field theory allows for the possibility of generating consciousness in the complete absence of neurons, using hardware instead of wetware to generate the relevant EM fields.

This proposed identity between consciousness and EM field patterns in principle provides a mechanism by which consciousness could initiate bodily movements. It well known that EM fields can influence and even cause neural activity.[1] Thus, at first sight it appears reasonable to suppose that conscious EM fields could act on the brain to cause bodily movements. This was initially seen (Pockett 2000; McFadden 2002a, 2000b) as a point in favour of the EM field theory of consciousness. However, doubts soon appeared (Pockett 2002) and have since hardened in the present author's mind to frank disbelief that conscious electromagnetic patterns do, in the normal course of events, cause bodily movements. The main reasons for this disbelief are as follows.

Reasons for Not Believing that Conscious Electromagnetic Patterns Cause Bodily Movements

1. Spatial EM patterns can only act on the neurons that generated them

The first reason for doubting that conscious EM patterns cause bodily movements concerns the physical characteristics of putatively conscious EM patterns. In order to understand the argument here, it is first necessary to understand the mechanism by which neurons generate the EM patterns in question.

The EM field patterns proposed by this theory as being conscious are spatial patterns of those transient, extracellular, electrical events known (quite independently of the present theory) as 'field potentials'. Field potentials are produced by the synchronous activation of chemical synapses on large numbers of anatomically aligned pyramidal cells in the cerebral cortex of the brain. The larger the number synapses on one pyramidal cell and/or the larger the number of spatially aligned post-synaptic pyramidal cells activated synchronously, the larger the field potential. The cellular mechanism by which

[1] See Adey 1981, Richardson *et al.* 1984, Taylor *et al.* 1984, Turner *et al.* 1984, Dalkara *et al.* 1986, Dudek *et al.* 1986, Snow and Dudek 1986, Yim *et al.* 1986, Faber and Korn 1989, Jeffereys 1995, Francis *et al.* 2003, Deans *et al.* 2007, Frölich and McCormick 2010.

field potentials are generated is well understood and is shown diagrammatically in *Figure 1*.

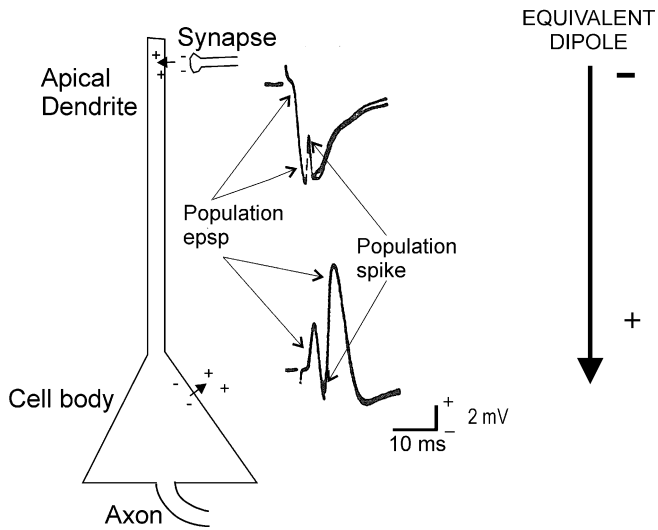


Figure 1: Diagram showing how the field potentials associated with a single pyramidal cell are generated. Activation of chemical synapses on the distal dendrites of pyramidal cells (top) causes a burst of positive ions to flow from the extracellular fluid into the apical dendrite. This leaves a short-lasting local negativity in the extracellular fluid around the synapse (top voltage trace). The product is a negative-going field potential known as a population epsp (excitatory post-synaptic potential). The intracellular positivity caused by this entry of ions to the dendrite essentially has to go somewhere, so to complete the circuit, positive ions flow out of the pyramidal cell at the cell body (bottom). This outflow produces a positive-going population epsp in the extracellular fluid near the cell body (bottom voltage trace). If the voltage transient in the cell body is large enough to cause an action potential to fire at the initial segment of the axon, a relatively small and brief "population spike" also appears in the middle of the population epsps.

Figure 1 illustrates two important points about the EM field theory of consciousness:

(a) The field potentials that make up putatively conscious fields are dominated by *synaptic potentials*, due to the activation of *chemical synapses*. The external fields produced by acton potentials *per se* have

relatively little influence on the shape of field potentials. The activation of electrical (as opposed to chemical) synapses has no influence on the shape of field potentials.

(b) The field potentials that contribute to conscious field patterns always come in positive-negative pairs: one positive-going field potential at the level of the pyramidal cell body and one negative-going field potential around the distal part of the apical dendrite. These positive-negative pairs are conveniently modelled as dipoles (right hand side of *Figure 1*). The EM field theory of consciousness proposes that conscious fields are distinguished from non-conscious fields – and from different conscious fields – by the spatio-temporal pattern in which these dipoles are arranged.

With this background, we may now return to the question of why it is doubtful that conscious EM fields can cause bodily movements. The exact features of the spatio-temporal dipole patterns that covary with particular sorts of consciousness are not yet clear (although some progress on that question is made by the arguments later in the present paper). However, one thing which is very clear is that such spatio-temporal patterns propagate very badly through space. Pockett *et al.* (2007) show mathematically that, assuming the patterns covarying with sensation to be produced by cortical dipoles spaced 3 mm apart, with an inter-pole distance of 2 mm (the approximate thickness of the cerebral cortex), the point spread function of EM fields through a medium with the conductivity of brain tissue is such that the patterns in question can not be recognised more than 2.5 mm above their source. In fact (i) some of the relevant dipoles may not extend the full width of the cortex, and (ii) the inter-dipole spacing is probably closer to 1 mm (the width of a human ocular dominance column) than 3 mm, which makes the situation even worse with respect to the distance over which a putatively conscious pattern could propagate – in any direction – and still be recognisable. In short, the physics of electromagnetism dictates that by the time one of these field patterns has travelled a mere mm or so from its source, it is so smeared or blurred as to be indistinguishable from a completely unpatterned (and therefore non-conscious) field.

This means that on any physically realistic version of the EM field theory of consciousness, individual conscious experiences remain very localized around the neurons that generate them. The only neurons a conscious field can

activate or influence *in a way that depends on the spatial pattern* of (and therefore the experience encoded by) *the field* are the very neurons that generated the field in the first place. In this regard, it is notable that all the empirical demonstrations cited earlier about the effect of brain-generated EM fields on the brain do involve the back-action of fields *on the neurons that generated them*.

How does this relate to the possibility of conscious fields' initiating bodily actions? It is true that the EM field theory of consciousness proposes that an individual's consciousness as a whole is the sum total of all the conscious fields generated by the brain. So in one sense, consciousness *as a whole* has access to the whole brain. But any particular conscious field pattern – for example, the conscious thought "I should check my mail-box within the next few minutes" – is able to affect the neuronal activity only of neurons in the immediate vicinity of those that generated the thought. A conscious thought such as "I should check the mail-box" would probably be generated somewhere in the prefrontal and/or parietal cortices (Pockett 2006). The main point of this argument is that it is physically impossible for a patterned EM field generated in the prefrontal and/or parietal cortex to have any direct effect on motor neurons, as suggested by McFadden (2002a, 2000b). Quite apart from the pattern-blurring effect of the point spread function, as mentioned above, the propagation of dipole electric fields obeys an inverse cube law. Motor neurons live in the spinal cord, several hundreds of mm away from either the parietal or prefrontal cortex. By the time it gets to the spinal cord, the EM field due to a parietal or prefrontal dipole has been reduced by the inverse cube law to less than 1/100,000 of its source strength. So never mind the fact that its spatial pattern would be unrecognisable – the EM field due to even the strongest parietal or prefrontal dipole essentially does not exist at all in the spinal cord.

Of course, a patterned EM field generated in the parietal and/or prefrontal cortex *could* have a direct effect on the neurons that generated it, and these neurons could then activate all the standard action potential and synaptic mechanisms by which the motor system is usually understood to work to produce an *eventual* effect on motor neurons. But the parietal and/or prefrontal neurons in question could just as easily do that job all by themselves, without the need for any intervention from a conscious EM field – or indeed without the need for any conscious thought at all. And oddly enough, the next section of the present paper cites considerable evidence that usually, this is exactly what happens.

2. Psychological evidence that consciousness does not cause behaviour

The second major reason for disbelieving that conscious EM patterns cause bodily movements concerns the psychologically measured properties of consciousness itself, independently of any theory as to its nature. The issue here is that there is now a large body of evidence that consciousness is *not* the direct cause of intentional aka voluntary movements.

Some of this evidence is summarised by Wegner (2002), Pockett (2004) and the various contributors to Pockett *et al.* (2006). The topic is too large allow its discussion in very great detail here, but in brief, consciousness is certainly not directly involved in the *control* of intentional movements. Jeannerod (2006) summarises a number of experiments showing that people are generally not even aware of having made the fine adjustments that serve to control their own intentional movements. The question of whether or not consciousness is involved in the *initiation* of voluntary movements is more complex, although ultimately the answer is just as clearcut.

The first and still most often quoted evidence against the idea that voluntary movements are initiated by consciousness is that generated by Libet *et al.* (1983). Libet and colleagues showed that the *Bereitschaftspotential* or readiness potential associated with a spontaneous voluntary movement begins some 350 ms before the subject reports having initiated the movement. This has been widely taken as demonstrating that voluntary movements are initiated un- or pre-consciously. Pockett and Purdy (2010) repudiate that conclusion by showing that:

(a) Readiness potentials are neither necessary nor sufficient (in anything other than a definitional sense) for voluntary movements. Rather, they simply indicate readiness to make a spontaneous movement.

(b) When subjects in Libet-style experiments are asked to make not spontaneous movements, but movements based on a definite decision about which of two acts to perform, the readiness potential is usually so much shorter than that associated with spontaneous movements that it starts at approximately the same time as the subject reports having initiated the movement.

These findings initially seem to reopen the possibility that voluntary movements might be initiated consciously. But the validity of *that* conclusion is immediately thrown into doubt by the subjective observation of one of Pockett

and Purdy's participants that for him, it was actually impossible to distinguish between deciding to move and moving. On closer inspection, this single observation turns out to be supported not only by the huge variability in the movement initiation times reported by the deliberately untrained participants in question (some of whom often reported that they had decided to initiate a particular movement after they had objectively moved), but also by a large body of experimental evidence from a whole series of other workers.[2] For example, Banks and Isham (2009) show that tricking a subject into thinking their own movements occurred progressively later than they actually did results in linear delays in the reported time of movement initiation.

The overwhelming conclusion from this body of experimental work is that humans do not consciously experience their decisions to initiate bodily movements in the same way as they experience sensory or even other cognitive events. Pockett and Miller (2007) show that subjects can report very accurately on the time at which they actually *make* a movement – perhaps because an actual movement is accompanied by a good deal of proprioceptive and other somatosensory feedback. But the psychological experiments of a number of labs now combine to suggest that when asked to report the time at which they initiated a movement, all a subject can do is infer that they must have initiated it sometime shortly before it took place.

The *coup de grace* in favour of this conclusion comes from the work of Desmurget *et al.* (2009), who describe a series of Penfield-esque experiments in which they directly stimulate various parts of the cerebral cortex in awake patients undergoing brain surgery and then ask the patients to report on their conscious experiences. When Desmurget and colleagues stimulate parietal regions, which are thought to generate conscious intentions to move at some time in the near future (Pockett 2006) the subject reports a subjective intention or desire to move the relevant part of the body, which escalates with stronger stimuli to a belief that they actually have moved, even though no EMG (electromyographic) activity can be detected. However, when the experimenters stimulate premotor cortex, which is thought to underpin the initiation of movements, the patient reports no subjective consciousness of movement and in fact firmly denies that any bodily movement has taken place,

[2] Aarts *et al.* 2005, Lau *et al.* 2007, Banks and Isham 2009, Kühn and Brass 2009, Rigoni *et al.* 2010.

even when the stimulation is turned up so much that the experimenters can *see* the relevant part of the body moving.

The addition of this neurophysiological evidence to the accumulated psychological results now makes it reasonable to conclude definitively that subjective reports about movement initiation are not reports of on-the-spot conscious perceptions at all. They are *post-hoc* cognitive constructions. We are not directly conscious of the initiation of our voluntary movements. Thus consciousness can not be considered to be the proximal cause of voluntary movements. The fact that experimental subjects usually fail to realise this is probably due to the general inability of humans to perceive cause and effect accurately (Choi and Scholl 2006).

Why is the Initiation of Movements not Accessible to Consciousness?

From the point of view of the EM field theory of consciousness, one beneficial effect of the fact that we do not consciously experience the initiation of bodily movements is that it renders harmless the parallel fact (see (1) above) that the sorts of EM fields the theory says are conscious can have direct physical effects only on the neurons that generated them. If consciousness itself does not initiate movements, there is no reason to suppose that putatively conscious EM fields should be able to initiate movements.

However, there is also another important consequence for the EM field theory of consciousness of the fact that we do not consciously experience the initiation of bodily movements. This is that it potentially provides a valuable clue as to the characteristics of dipole patterns that *do* underpin conscious experiences.

As explained earlier, the EM field theory of consciousness proposes that different putatively conscious EM fields are characterised by different spatial dispositions of the standard 'field potentials' generated by activation of chemical synapses on the distal apical dendrites of cortical pyramidal cells (*Figure 1*). The spatial disposition of field potentials clearly depends on the anatomical disposition of pyramidal cells. This means that for the EM field theory of consciousness to work:

(a) parts of cortex that produce different *types* of conscious experience should probably have different types of intracortical disposition of pyramidal cells

(b) parts of the cortex that do not produce conscious experiences at all should definitely have a fundamentally different intracortical disposition of pyramidal cells from parts of the cortex that do produce conscious experiences.

Fortunately for the theory, it turns out that requirement (a) was shown to be broadly fulfilled a little over a hundred years ago. *Figure 2* reproduces the map of spatial variations in the cytoarchitecture of the cerebral cortex published by Brodmann (1909).[3] This map is based on microscopically observed differences in the layering of pyramidal and other neurons from the outside to the inside of the cortex. Despite a mid-century flurry of complaints from Lashley (who, unable to find his 'engram', became convinced that the cortex must be pluripotential in function and therefore uniform in structure), the cytoarchitectonic areas delineated by Brodmann have since been shown to correspond so well with the functional areas delineated by modern electrophysiological and imaging techniques that particular regions of cortex are still routinely identified using their Brodmann Area (BA) numbers. For our present purposes the most relevant of these numbers are those of the motor and pre-motor cortices (BA4 and BA6).

[3] Although Brodman's map is almost identical to an earlier map by Campbell (1905), the Brodmann version has become much more well-known, to the extent that its numbering system is still widely used.

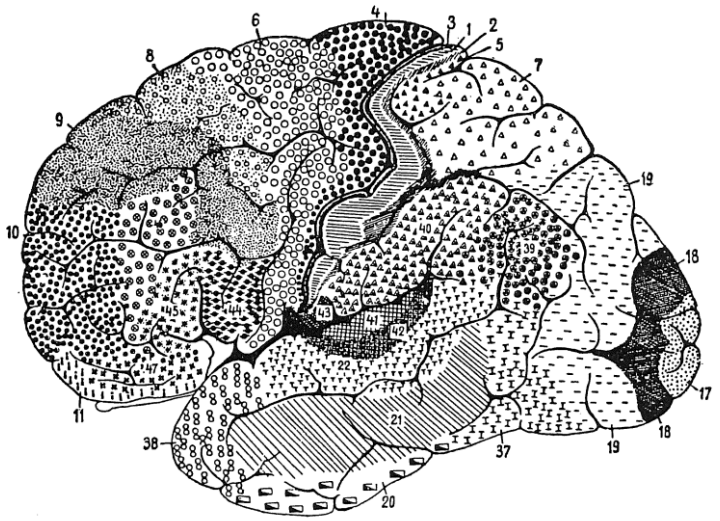


Figure 2: Brodmann's cytoarchitectural map of the human cerebral cortex. Brodmann Areas 4 & 6 are motor & premotor cortex. Brodmann Areas 1,2 & 3 (somatosensory), 17 & 18 (visual) and 41 & 42 (auditory) are primary and sccondary sensory cortex. Cerebellum not shown.

So requirement (a) above has been known for a century to be largely fulfilled. One very useful effect of the new information that movement initiation is inaccessible to consciousness is that it allows us now to test requirement (b).

While the longer term *planning* of movements (which is accessible to consciousness) probably occurs in the prefrontal and/or parietal cortices (BA 10, 11, 39 & 40), the *initiation* of movements is generally thought to occur somewhere in BA6, the area now known as pre-motor cortex (see Pockett 2006 for review). It is therefore deeply fortunate for the EM field theory of consciousness that the anatomical work of Brodmann's successor von Economo (1925 and 1927) does indeed show a fundamental difference between the cytoarchitectonics of premotor and motor cortices (BA6 & BA4) and that of all other regions of the neocortex.

Figure 3 shows in diagrammatic form the cortical locations and cell body architectures of the five general types of cortex identified by von Economo. *Figure 3(A)* shows that the area encompassed by BA6 and BA4 is classified as 'agranular' cortex. *Figure 3(B)* shows that the major difference between agranular cortex and the other four cortical types in Economo's classification

system is that agranular cortex lacks the 'granular' layer which in most cytoarchitectonic classification systems is called lamina IV.

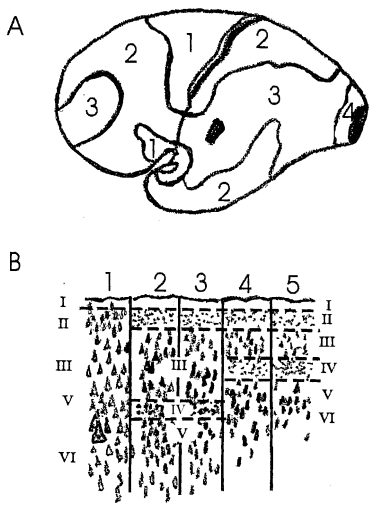


Figure 3: Cortical locations (A) and cellular structure (B) of the cytoarchitectonic types classified by von Economo (1925 and 1927). Roman numerals in B indicate cytoarchitectonic layers: surface of brain at top. Lamina I consists entirely of apical dendrites from the pyramidal cells of lower layers (see *Fig 4*) and incoming axons from other parts of the cortex, which synapse on the dendrites. Apart from lamina I, cytoarchitectonic type 1 (found mainly in areas BA 4 & 6) lacks well-demarcated layers. The five architectonic types named by von Economo are 1- agranular cortex; 2 - frontal cortex; 3 – parietal cortex; 4-polar cortex; 5 (solid black areas in A) – granular or koniocortex.

In all regions of the neocortex except the motor and premotor cortices, lamina IV is a largely pyramid-free layer of stellate neurons, which in Nissl stained sections[4] somewhat resemble dust (leading to von Economo's identification of the three major primary sensory areas, where lamina IV is particularly obvious, as 'koniocortex'). The stellate neurons in Lamina IV

[4] Nissl staining shows all cell bodies, but none of their axons or dendrites. Golgi staining selects a few neurons (in a completely uncontrollable way) and stains the whole cell, including the axon and all dendrites.

receive major synaptic input from the subcortical thalamus. However the contribution of these thalamo-cortical synapses to the field potential landscape would be negligible, because the dendrites of stellate cells extend at all angles from the cell body and thus the positive and negative voltage transients illustrated in *Figure 1* tend to cancel each other out. In contrast, the apical dendrites of most of the pyramidal cells in laminae III and IV – including those in motor and pre-motor cortices (Porter and Lemon 1993) – extend all the way up to lamina I, at the surface of the cortex (see *Figure 4*).

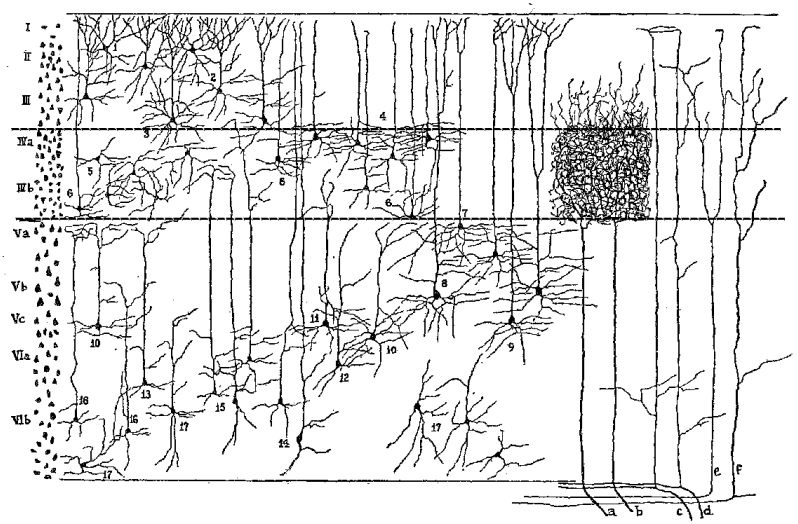


Figure 4: Semi-diagrammatic view of Golgi stained neurons in adult mouse parietal cortex (after Lorente de No, 1951). Axons are not drawn in the original "to avoid complication". Nissl-stained cells at left illustrate cortical layers. Dotted lines added by present author. Differences in detail doubtless exist between mouse & human and between different cortical areas within a given species: this figure is intended simply to illustrate general structural features of non-motor cortex.

In terms of the overall disposition of dipoles, this tendency for the apical dendrites of most pyramidal cells to extend right up to the surface of the cortex means that those parts of the neocortex which do generate conscious experiences are likely to produce a field potential landscape characterised by one layer of negative poles at the surface of the brain and two deeper layers of positive poles, separated by an electrically neutral field in the position of

lamina IV (*Figure 5*). In contrast, any area of brain that does not produce conscious experiences is likely to produce something more like a simple surface-negative, depth-positive dipole, without the complex double-layering of the deeper positive field. This latter class of brain areas would definitely include the cerebellum, which (a) does not generate conscious experiences (Jeannerod 2006) and (b) does not exhibit anything like the 6-layered anatomical structure characteristic of neocortex (Eccles *et al.* 1967).

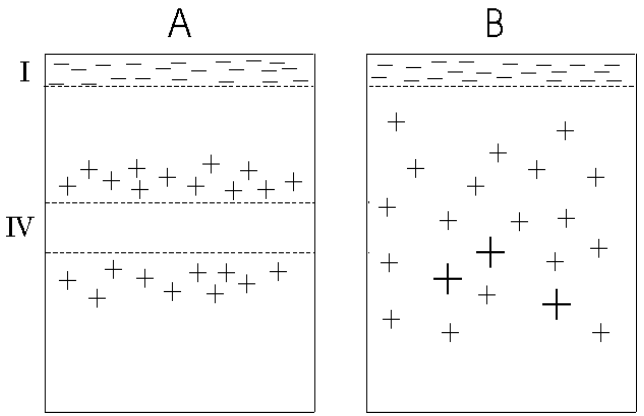


Figure 5: Schematic diagram of the proposed EM field shapes characterising conscious (A) and unconscious (B) fields. Roman numerals on left of diagram indicate cortical layers: Lamina I is at the surface of the cortex. The sprinkling of larger positive charges in B is intended to represent the giant Betz cells found in motor cortex.

For logistical reasons it would not be particularly easy (although not impossible) to perform experimental tests in human subjects of these predictions about the dipole landscapes produced by parts of the cortex that do and do not generate conscious experiences. But the predictions made here could relatively easily be tested in animals. Pending such testing, these speculations may point the way with regard to further elaboration of the EM field theory of consciousness. However, considerably more work on the relationship between cytoarchitecture and neurophysiological activity is necessary before any more detailed predictions can be made about the EM field shapes underpinning the various different modalities of conscious experience.

REFERENCES

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition, 14*(3), 439-458.

Adey, W. R. (1981). Tissue interactions with nonionizing electromagnetic fields. *Physiological Reviews*, *61*(2), 435-514.

Banks, W. P., & Isham, E. A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science, 20*(1), 17-21.

Brodmann, K. (1909). Vergleichende Lokalisationslehre der Grosshirnrinde: in ihren Prinzipien dargestellt auf Grund des Zellenbaues. Leipzig: Verlag von Johannes Barth. [*Brodmann's localisation in the cerebral cortex: the principles of comparative localisation in the cerebral cortex based on cytoarchitectonics*. Tr. by J. L. Garey. Springer, 1994]

Campbell, A. W. (1905). *Histological Studies on the Localization of Cerebral Function*. Cambridge: Cambridge University Press.

Choi, H., & Scholl, B. J. (2006). Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception. *Perception, 35*, 385-399.

Dalkara, T., Krnjevic, K., Ropert, N., & Yim, C. Y. (1986). Chemical modulation of ephaptic interaction of CA3 hippocampal pyramids. *Neuroscience*, *17*(2), 361-370.

Deans, J. K., Powell, A. D., & Jefferys, J. G. (2007). Sensitivity of coherent oscillations in rat hippocampus to AC electric fields. *J. Physiol., 583*, 555-565.

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement intention after parietal cortex stimulation in humans. *Science, 324*, 811-813.

Dudek, F. E., Snow, R. W., & Taylor, C. P. (1986). Role of electrical interactions in synchronization of epileptiform bursts. *Advances in Neurology, 44*, 593-617.

Eccles, J. C., Ito, M., & Szentagothai, J. (1967). *The cerebellum as a neuronal machine*. New York: Springer.

Economo, C. von (1925). Die fünf Bautypen der Grosshirnrinde. *Schweiz Arch Neurol Psychiatri (Zürich), 16*, 266-269.

Economo, C. von (1927). *Zellaufbau der Grosshirnrinde des Menschen*. Aehn Vorlesungen. Verlag von Julius Springer.

Faber, D. S., & Korn, H. (1989). Electrical field effects: Their relevance in central neural networks. *Physiological Reviews, 69*(3), 821-863.

Feigl, H. (1958). The 'mental' and the 'physical'. In H. Feigl, G. Maxwell & M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science,* Vol 2. Minneapolis: University of Minnesota Press.

Francis, J. T., Gluckman, B. J., & Schiff, S. J. (2003). Sensitivity of neurons to weak electric fields. *Journal of Neuroscience, 23*(19), 7255-7261.

Frölich, F., & McCormick, D. (2010). Endogenous electric fields may guide neocortical activity. *Neuron, 67*(1), 129-143.

Jeannerod, M. (2006). Consciousness of action as an embodied consciousness. In S. Pockett, W. Banks & S. Gallagher (Eds.), *Does Consciousness Cause Behavior?* (pp. 25-38). Cambridge, MA: MIT Press.

Jefferys, J. G. (1995). Nonsynaptic modulation of neuronal activity in the brain: Electric currents and extracellular ions. *Physiological Reviews, 75*(4), 689-723.

Kühn, S., & Brass, M. (2009). Retrospective construction of the judgment of free choice. C*onsciousness and Cognition, 18*, 12-21.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience, 19*(1), 81-90.

Libet, B., Gleason C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain, 106*(3), 623-642.

Lorente de No, R. (1951). Cerebral cortex: Architecture, intracortical connections, motor projections. In J. F. Fulton (Ed.), *Physiology of the Nervous System*, (pp. 288-330). New York: Oxford University Press.

McFadden, J. (2002a). Synchronous firing and its influence on the brain's electromagnetic field: Evidence for an electromagnetic field theory of consciousness. *Journal of Consciousness Studies, 9*(4), 23-50.

McFadden, J. (2002b). The conscious electromagnetic information (cemi) field theory: The hard problem made easy? *Journal of Consciousness Studies, 9*(8), 45-60.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*(1): 44-50.

Pockett, S. (1999). Anesthesia and the electrophysiology of auditory consciousness. *Consciousness and Cognition, 8*(1), 45-61.

Pockett, S. (2000). *The Nature of Consciousness: A Hypothesis*. Lincoln, NE: iUniverse.

Pockett, S. (2002). Difficulties with the electromagnetic field theory of consciousness. *Journal of Consciousness Studies, 9*(4), 51-56.

Pockett, S. (2004). Does consciousness cause behaviour? *Journal of Consciousness Studies, 11*(2), 23-40.

Pockett, S. (2006). The neuroscience of movement. In S. Pockett, W. Banks & S. Gallagher (Eds.), *Does Consciousness Cause Behavior?* (pp. 9-24). Cambridge, MA: MIT Press.

Pockett, S., Banks, W., & Gallagher, S. (Eds.) (2009). *Does Consciousness Cause Behavior?* Cambridge, MA: MIT Press. [2006]

Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition, 16*(2), 241-254.

Pockett, S., & Purdy, S. (2010). Are voluntary movements initiated preconsciously? The relationships between readiness potentials, urges and decisions. In W. Sinnot-Armstrong & L. Nadel (Eds.), *Conscious Will and Responsibility: A Tribute to Benjamin Libet.* New York: Oxford University Press.

Pockett, S., Zhou, Z. Z., Brennan, B. J., & Bold, G. E. J. (2007). Spatial resolution and the neural correlates of sensory experience. *Brain Topography, 20*(1), 1-6.

Porter, R., & Lemon, R. (1993). Corticospinal function and voluntary movement. *Monographs of the Physiological Society, 45*. Oxford: Clarendon Press.

Richardson, T. L., Turner, R. W., & Miller, J. J. (1984). Extracellular fields influence transmembrane potentials and synchronization of hippocampal neuronal activity. *Brain Research, 294*(2), 255-262.

Rigoni, D., Brass, M., & Sartori, G. (2010). Post-action determinants of the reported time of conscious intentions. *Frontiers in Human Neuroscience,* 4(38).
Published Online, doi: 10.3389/fnhum.2010.00038.

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review, 68*(2), 141-156.

Snow, R. W., & Dudek, F. E. (1986). Evidence for neuronal interactions by electrical field effects in the CA3 and dentate regions of rat hippocampal slices. *Brain Research, 367*(1-2), 292-295.

Taylor, C. P., Krnjevic, K., & Ropert, N. (1984). Facilitation of hippocampal CA3 pyramidal cell firing by electrical fields generated antidromically. *Neuroscience, 11*(1), 101-109.

Turner, R. W., Richardson, T. L., & Miller, J. J. (1984). Ephaptic interactions contribute to paired pulse and frequency potentiation of hippocampal field potentials. *Experimental Brain Research, 54*(3), 567-570.

Wegner, D. M. (2002). *The Illusion of Conscious Will.* Cambridge, MA: MIT Press.

Yim, C. C., Krnjevic, K., & Dalkara, T. (1986). Ephaptically generated potentials in CA1 neurons of rat's hippocampus in situ. *Journal of Neurophysiology, 56*(1), 99-122.

The Dynamics of Acting

Mark H. Bickhard *
mhb0@lehigh.edu

ABSTRACT

A model of decision and action processes is outlined, and several consequences of this model are developed. The simple plausibility of this model demonstrates that common discussions of decision making and action are constrained within metaphysical frameworks that are at best questionable. The model, in turn, enables a model of free will that is consistent with contemporary physics and likely, from an evolutionary perspective, to have emerged. The model is an alternative to decision making as computation and action as the initiation of a causal chain, and entails that decision and action processes are self-organizing, global, and intrinsically intertwined with each other. In turn, the model broadens attention with regard to ethical considerations to persons and their character, and to the development and developing processes of those persons and characters. This is much closer to an Aristotelian virtue ethics than to a primarily action focused ethics, such as Kant's. Finally, the overall model of persons that emerges in the discussion is one of ongoingly active and ongoingly developing processes, not one of some kind of classical substance or entity.

---

1. GLOBAL RELATIONALISM

Action and acting are commonly investigated within a framework assumption that action is a special kind of, or a special initiation of, causal chains. Causal chains, in turn, are chains of particular event-points through which some kind of causal influence proceeds. In concert, decision making is commonly investigated as a special kind, or a special result, of computational reasoning. A decision, thus, would typically be a computation that terminates with the initiation of a causal chain.

* Department of Philosophy and Department of Psychology – Lehigh University (Bethlehem, PA - USA)

I contend that both assumptions are in error, and that they yield serious aporia concerning the nature of decision making and acting, with resultant distortions in related investigations, such as of free will and responsibility. I outline an alternative model that avoids these errors and distortions.

Arguments against causal chain models of action and computationalist models of decision making have been elaborated elsewhere (e.g., Bickhard 1996, 2009a, in press-a, in preparation), so I will just outline some of the central points here. Regarding causal chain models: there is an underlying metaphysical assumption in causal chain models that causality can somehow travel from particular point-event to particular point-event, which, in turn, presupposes that our world consists of such point-events. But there are reasons to reject this (Bickhard, in press-a, in press-b; Butterfield 2006). One is that a continuum, such as that of space-time, cannot be constructed out of particular points. A related reason is that process models, such as those of quantum field theory, force a relationalism, and are not consistent with particularisms. But, if particular point-events cannot constitute the underlying furniture of our world, then real processes cannot be constituted out of (causal chains of) such point-events. Real processes in the world are extended in both space and time, and can have global properties that cannot be captured with causal chains. This is of particular importance for later discussion in this article with regard to self-organizing processes.

Regarding computational models: such models assume that representation is constituted as some form of encoding of what is represented, but such encodingism assumptions are, at root, incoherent (Bickhard 1996, 2009a, 2009b, in preparation; Bickhard and Terveen 1995). But, if computationalist style representations cannot be the fundamental form of representation, then computation on such "symbolic" representations cannot be the fundamental form of thought or decision making. Connectionist models might at first seem to be immune to such "encodingist" critiques, but presumed connectionist representations are still "just" trained encoding correspondences, and do not avoid the basic problems involved.

If so, then alternatives must be found. I turn now to the general form of such an alternative, one that integrates issues of decision making and issues of acting.

2. Decisioning and Acting

The general model framework that I will be making use of here is that of self-organizing processes. In general, decision processes are constraint-satisfying self-organizing processes, and acting is *temporally extended* constraint satisfying self-organizing processes. In this paper, I will not be arguing directly for these models, though I will give some indications of their plausibility (see Bickhard, in press-a, in preparation; Juarrero 1999). The central point here is that such models *are* plausible, and that they are inconsistent with the presuppositions of standard models, and that they therefore yield (plausible) alternatives to common forms of argument and conclusions concerning such issues as free will. Arguments that such models are correct, as well as plausible, require much more extensive development (Bickhard, in preparation).

A first shift in perspective that is required to understand these kinds of models is to recognize that the central nervous system, from single neurons to the entire system (and, in fact, including the entire organism) is *always* active. It is always doing something; to do nothing is to be dead. Thus, neurons are not passive threshold switches and the CNS is not a passive information processor. Functional relationships with the world, thus, as well as relationships within the CNS, are not those of transduction, triggering, or equivalent activities on a passive system, but, instead, are those of the modulations of intrinsically already ongoing activity, of the organism, of the CNS, of domains of CNS activity, of single neurons, of astrocytes, etc. (Bickhard 2008a).

Such ongoing activity in the nervous system is not totally free: it is functional. Such activity functions to modulate (control is a strong version of modulation) other activity, including that of acting in the world. The crucial point here is that this is a normative functionality: such activities and the modulations that they induce can be inappropriate to the environment (for example). They can be dysfunctional.

But that entails that the ongoing activity at a given time constitutes a preparation, a set-up, for the functional activity, e.g., action and interaction, that will ensue. Just as that activity can be inappropriate or wrong, so can the preparations for that activity: there is an anticipatory character to how the system is prepared and preparing to function, and such anticipation can be false.

This is the core of an action and interaction based model of representation. Truth value is the central problem of representation, and anticipation yields emergent truth value.[1] More familiar kinds of representation, such as of objects, can be constructed out of such "representations as anticipations" in a roughly Piagetian manner (Bickhard 2009a, in preparation; Piaget 1954).

2.1. Constraint satisfying self-organizing processes

Consider now the possibility that some CNS activities can serve as (soft) constraints on other CNS activities. The constraining activities could effect such constraints via a form of modulation of other self-organizing processes, such that the self-organizing processes would be modulated to honor those constraints. If the constraining processes were to constitute representational conditions, then this would constitute a model of internal constraint satisfaction problem solving, with the constraining processes constituting the problem "definitions".

In turn, this would constitute a model of an internalization of a variation and selection process – an evolutionary epistemology (Campbell 1974). The constraints serve as selection principles and the self-organizing processes generate potential satisfiers of those selection principles.[2]

Crucially, such self-organizational processes constituting evolutionary epistemological processes would be intrinsically global in nature, and not capturable with point to point causal event chains.

2.2. Decision processes

A self-organizing evolutionary epistemological model can also serve as a model of decision processes. A decision involves determining what to do, given some constraints on what is desirable and permitted. In general, these constraints will involve both environmental conditions, and internal criteria such as preferences, goals, and values. Taking those external and internal conditions as constraints on a problem solving variation and selection process yields a candidate for satisfying those constraints– yields a (candidate) decision "outcome". Note that this is not at all a computational model of decision making.

[1] For a model of ontological emergence, see Bickhard 2009a.

[2] See Bickhard 2002, 2009a for a model of rational thought based on these notions.

2.3. (INTER)ACTION PROCESSES

CNS processes may or may not modulate activity that controls muscles. If so, we call the overall process a process of (inter)action. There is no constraint in the model of constraint satisfying self-organization outlined above on whether or not the activities do, in fact, generate action; they may or they may not, but the general self-organizing nature of the processes will be the same in either case. That is, the general model can serve also as a model of interaction, with the emergent constraint satisfying processes modulating downstream muscle activity.

2.4. GLOBAL AND INTERTWINED

Such a model of acting is not consistent with action as initiated at some event-point, which then pursues a causal-chain trajectory through further such event-points. Such a self-organizing model of acting is both global, in the sense that the processes are determined (if at all) by global characteristics, not reducible to local causal chains, and extended, in the sense that acting is an ongoing process of modulation and control, extended in both space and time, with feedback, monitoring, and adjustments. Acting does of course have (modulatory) consequences beyond the range of such feedback, etc., – e.g., the thrown rock hits something, makes a sound, etc. – but 1) those too will not in general be reducible to causal chains, and 2) such further consequences can be descriptively part of an "action", perhaps even in the sense that they are represented in the goals, etc. that constrain the internal processes, but the fact that some consequences of acting are beyond the range of feedback and other internal forms of self-organization and its modulation does not entail that acting in general is of that form.

In fact, it is clear that acting in general cannot be in that form. Acting is in accordance with preferences, goals, and values in terms of (heuristic) strategies, sub-strategies, alternative strategies, subsidiary further problem solving, etc. that constitute a potential "fit" to those constraints. Plans are never detailed down to the most minute sub-action; they are always at more general levels, which have to be "filled-in" progressively in ongoing interacting. Acting is globally organized (and organizing); it is extended and organized(ing) in space and time.

Notice that some of these aspects of acting are themselves constitutive of decision processes – e.g., a decision about a sub-goal or strategy. Thus, both

decision processes and acting processes are global and extended, *and* they are inherently intertwined. The "picture" of computational decision processes that end with the initiation of an action in the form of a causal chain is incorrect even at a general descriptive level: the processes involved are not point-events, and the two kinds of processes are not dynamically distinct. Decision-acting processes are global and extended and are distinguishable *aspects* of a single underlying kind of process. They are not pointillistic.

3. Free Will

Free will has to do with the unpredictability and indeterminancy of decision and action processes. I will argue here that free will in those senses is possible, and perhaps even likely, but, in the next section, that it doesn't make much difference for issues of moral responsibility and evaluation.

3.1. Unpredictability

There are circumstances in which it can be adaptive to be unpredictable in one's actions. This is particularly clear in conditions of conflict: if an opponent can predict what you will do, then the opponent may have an advantage. There are reasons to think that social phenomena such as this were strongly involved in the evolution of the human brain (Humphrey 1976; Byrne and Whiten 1988), and, thus, that the adaptiveness of unpredictability was similarly a strong selection constraint. The general point, however, is much broader than that: it holds, for example, in most predator-prey relationships.

Selection for the possibility of unpredictability can be "easily" satisfied: chaotic processes are intrinsically unpredictable. Chaotic processes are highly sensitive to initial conditions, so much so that those initial conditions cannot be determined with sufficient accuracy to permit prediction. Chaotic processes are, nevertheless, deterministic processes. So, chaos generates a differentiation between predictability and determinancy (Bickhard, in press-b). If the processes that generate candidate satisfiers of internal problem defining constraints are themselves chaotic, then we have a model of unpredictable, even though determinate, decision-acting processes.

3.2. Indeterminism

Selection for unpredictability may, then, constitute selection for chaotic processes, which are chaotic in virtue of high sensitivity to initial conditions. If such sensitivity to initial conditions were of sufficiently fine scale, then it could include sensitivity to conditions at a quantum level. Quantum level conditions involve an inherent indeterminancy, and, thus, such a process would be not only unpredictable, but also indeterministic. There may be a major metaphysical distinction between a process being unpredictable but deterministic and a process being intrinsically indeterminate, but that distinction is likely not "visible" to selection constraints for chaotic processes. So evolutionary selection for unpredictability may well have generated indeterministic sensitivity to quantum level conditions.

Such recruitment of inherently indeterministic quantum level phenomena into global decision and action processes, therefore, will constitute a model of decision and acting processes that are both unpredictable and indeterministic. This would be an "amplification" of quantum level indeterminancy to the level of decision and action (Bishop 2002).

Inherent unpredictability and indeterminancy of decision and action processes constitutes a good candidate model for free will. It at least captures the primary characteristics commonly taken to constitute free will. So, given the possibility of high quantum level sensitivity of "chaotic" processes, free will is at least possible. The evolutionary considerations suggest that it might even be likely.

4. Ethical Considerations

Free will is of relevance most centrally to issues of ethics and morality. So a clear next question is what relevance this model of free will has for ethical and moral issues.

4.1. Responsibility

A model of pure random action is not a model of free will (Dennett 1984). One aspect of this point is that, in such a model, there is no person who is engaged

in and modulatory of such action.[3] Free will, if it exists at all, must be a property of a person's activities: "simple" randomness does not suffice.

In the model outlined, the randomness, thus unpredictability and indeterminism, is within the constraints of preferences, goals, values, environmental considerations, and so on. The ongoing decision-action processes will, in general, honor these constraints, and, thus, indirectly manifest them. These normative constraints constitute core aspects of persons; they constitute core aspects and parts of character. Activity, then, will tend to honor and be manifestations of the (character of the) person engaged in that activity.

At this point, the globality of self-organization has an important consequence: self-organizing processes cannot be reduced to causal chains, and, especially if they are at least chaotic, cannot be canceled out in favor of prior causal chains leading into self-organization. External constraints and conditions can certainly influence decision and action process, and can certainly influence the construction of preferences, goals, values, and other normative aspects of character, but they cannot determine them. Similarly, the quantum randomness of constructions that occur *within* such constraints can influence later aspects of character, but they cannot determine them.

The general point here is that it is not possible to cancel out the character of a person in favor of prior "luck" in the form of prior causal chains. A person and their character is ineliminable in accounting for decision and action processes. Whatever the influences may be of incoming "causality", it cannot be exhaustive. Consequently, a person and their character is ineliminable in considerations of responsibility for decision and action. Ethical evaluations, then, will necessarily involve evaluations of persons.

Note that this point does not depend on in-principle indeterminate free will. The *character* of persons constitutes the primary locus of ethical ontology, no matter the determinancy or indeterminancy of the self-organizing processes of decisions and actions. So, free will may be possible and even likely, but, I contend, it doesn't make much difference with regard to ethical issues.

4.2. Ethics

I have outlined a model of ethical ontology that focuses on persons and their character. This is much closer to an Aristotelian framework than to a Kantian

[3] For a model of persons consistent with the above discussion, see Bickhard 2008 , in press-c.

framework, and it is worth pointing out some of the differences that follow from this primary distinction.

One of the most important is that such a model of acting as honoring character shifts the focus of ethical ontology form that of action to that of character – toward a virtue ethics. But the difference is even stronger than this. Action honors character, but character is itself developed, and, furthermore, character is in ongoing development throughout a person's life time (Bickhard 2006, 2008b, in press-c).

There are, thus, at least three levels of person-level ethical ontological consideration: "selecting actions, selecting kinds of person to become, selecting kinds of becoming to engage in" (Bickhard, in press-a). A strong, or exclusive, focus on action as the locus of ethics obscures considerations of learning and development, and of the ethical issues involved in processes of learning and development. An action focus obscures the possibilities of ethical error in a person's being, and in a person's becoming.

5. Conclusions

I have limned a model of decision and action processes that constitutes an alternative to models that are embedded in assumptions of point-event causal chains. By taking into account the anticipatory inherent activity of the organism and nervous system, non-computational models of mental process can be developed, and non-causal chain models of decision and action processes. I have not argued in detail for these models: their very existence, so long as they are at all plausible, suffices to show that much of the current literature proceeds within questionable presupposed frameworks.

Some further consequences of such models are: 1) a model of free will that is consistent with contemporary physics, and at least likely from an evolutionary perspective, and 2) an ontology for ethics that emphasizes persons and their character, not just actions, and, furthermore, illuminates the ethical relevance of the *development* of persons, not just the actions that they engage in. Persons are not things or entities or substances: persons are open, ongoingly self-organizing, at multiple temporal scales, processes.

REFERENCES

Bickhard, M. H. (1996). Troubles with computationalism. In W. O'Donohue & R. F. Kitchener (Eds.), *The Philosophy of Psychology,* (pp. 173-183). London: Sage.

Bickhard, M. H. (2002). Critical principles: On the negative side of rationality. *New Ideas in Psychology, 20*(1), 1-34.

Bickhard, M. H. (2006). Developmental normativity and normative development. In L. Smith & J. Voneche (Eds.), *Norms in Human Development*, (pp. 57-76). Cambridge: Cambridge University Press.

Bickhard, M. H. (2008a). *The Microgenetic Dynamics of Cortical Attractor Landscapes.* May 22-23, 2008. Workshop on "Dynamics in and of Attractor Landscapes". Isola d'Elba, Italy: Parmenides Foundation.

Bickhard, M. H. (2008b). Are you social? The ontological and developmental emergence of the person. In U. Müller, J. I. M. Carpendale, N. Budwig & B. Sokol (Eds.), *Social Life and Social Knowledge*, (pp. 17-42). New York: Taylor & Francis.

Bickhard, M. H. (2009a). The interactivist model. *Synthese*, *166*(3), 547-591.

Bickhard, M. H. (2009b). Interactivism. In J. Symons & P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology*, (pp. 346-359). London: Routledge.

Bickhard, M. H. (in press-a). Some consequences (and enablings) of process metaphysics. *Axiomathes.*

Bickhard, M. H. (in press-b). Systems and process metaphysics. In C. Hooker (Ed.), *Handbook of Philosophy of Science. Philosophy of Complex Systems*, (Vol. 10). Amsterdam: Elsevier.

Bickhard, M. H. (in press-c). A process ontology for persons and their development. *New Ideas in Psychology.*

Bickhard, M. H. (in preparation). *The Whole Person: Toward a Naturalism of Persons – Contributions to an Ontological Psychology.*

Bickhard, M. H., & Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution.* Amsterdam: Elsevier Scientific.

Bishop, R. C. (2002). Chaos, indeterminism, and free will. In R. Kane (Ed.), *The Oxford Handbook of Free Will*, (pp. 111-124). Oxford: Oxford University Press.

Butterfield, J. (2006). Against Pointillisme about mechanics. *British Journal for the Philosophy of Science, 57*(4), 709-753.

Byrne, R. W., & Whiten, A. (1988). *Machiavellian Intelligence*. Oxford: Oxford University Press.

Campbell, D. T. (1974). Evolutionary Epistemology. In P. A. Schilpp (Ed.), *The Philosophy of Karl Popper*, (pp.413-463). LaSalle, IL: Open Court.

Dennett, D. C. (1984). *Elbow Room*. Cambridge, MA: MIT Press.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing Points in Ethology*, (pp. 303-317). London: Cambridge University Press.

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: MIT Press.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.

Deliberative Libertarianism*

Jing Zhu **
zhujing6@mail.sysu.edu.cn

ABSTRACT

Libertarian theories of free will maintain that the freedom of will is incompatible with determinism, and that the involvement of indeterminacy can somehow lead to genuine free actions, in the sense that an agent has a choice about what to do, is able to do other than what she actually does, and is the ultimate source of her responsible actions.[1] Philosophers disagree on where the indeterminism required for libertarian free agency is located in the processes leading to an agent's practical decision and rational action. Deliberative indeterminism or libertarianism holds that the required indeterminism should take place relatively early in the process of deliberation[2], prior to the momentary mental act of decision-making that terminates a deliberation.[3] Ironically, the idea of deliberative libertarianism has been proposed and developed mainly by some non-libertarians, most notably by Daniel Dennett and Alfred Mele[4], and is opposed by some libertarians.[5] The

* The author is grateful to Alfred Mele and Robert Kane for valuable comments on a draft of this article. This research was support by a grant (06BZX024) from National Social Sciences Foundation of China.

** Institute of Logic and Cognition – Department of Philosophy – Sun Yat-sen University (Guangzhou, China); Institute of Philosophy of Mind and Cognition – National Yang-Ming University (Taipei, Taiwan)

[1] See Kane 1996 and Clarke 2003 for recent reviews of libertarian accounts of free will.

[2] The terms "deliberative indeterminism" and "deliberative libertarianism" are adopted from Clarke 2002 and 2003 respectively.

[3] See Mele 2003, ch. 9 for an articulation and defense of the idea that to decide to do something is to perform a momentary mental action of forming an intention to do it, which terminates the agent's practical deliberation.

[4] Dennett is a well-known compatibilist, and Mele is a self-claimed agnostic about the truth of compatibilism.

[5] Robert Kane, a leading contemporary libertarian, once proposed a richer account of deliberative libertarianism than Dennett's in his *Free Will and Values* (1985), which contains many valuable germinal ideas to be developed in this paper. However, it seems that Kane tries to distance himself from this position lately by declaring that he believes that the idea of deliberative libertarianism is only a part but «not adequate in itself even for an account of free practical choice» (Kane 1996, pp. 162 and 236) and that he has never unqualifiedly endorsed it (2002, p. 25). In his recent *A*

aim of this paper is to develop and defend deliberative libertarianism. Section 1 describes the basic idea of deliberative libertarianism. Section 2 surveys some major objections directed against it. I defend deliberative libertarianism in Section 3 after refining the psychological model of deliberation and decision-making proposed by this brand of indeterminism, and conclude with some suggestions for libertarianism in general in Section 4.

1. Deliberative Libertarian Accounts

A persistent charge against libertarianism is that, even if determinism is false, the involvement of indeterminacy, which implicates randomness, chanciness and arbitrariness, can still hardly secure a condition for rational, responsible free action. Any positive libertarian theory of free will is faced with the challenge to provide an intelligible, coherent and plausible account on how a certain kind of indeterminism can be freedom-enhancing, rather than freedom-diminishing, in the production of an agent's rational and responsible action. The idea of deliberative libertarianism has hence been proposed and recommended to libertarians to cope with this challenge.

1.1. Dennett

In "On Giving Libertarians What They Say They Want" (Dennett 1978), Dennett suggests that it may be possible to install indeterminism at the microscopic level in the internal causal chains that affect human behavior at the macroscopic level while preserving the intelligibility of practical deliberation that libertarians require (Dennett 1978, pp. 290-292), and that the required indeterminism should be placed «at some earlier point, prior to the ultimate decision or formation of intention» (Dennett 1978, pp. 292-293). Dennett goes on to propose the following "realistic model of decision-making":

> When someone is faced with an important decision, something in him generates a variety of more or less relevant considerations bearing on the decision. Some of these considerations, we may suppose, are determined to be generated, but others may be non-deterministically generated. For instance,

Contemporary Introduction to Free Will (2005), while surveying Dennett's and Mele's views under the section title "Deliberation and Causal Indeterminism" (pp. 64-65), Kane does not mention his own early work on this topic.

> Jones, who is finishing her dissertation on Aristotle and the practical syllogism, must decide within a week whether to accept the assistant professorship at the University of Chicago, or the assistant professorship at Swarthmore. She considers the difference in salaries, the probable quality of the students, the quality of her colleagues, the teaching load, the location of the schools, and so forth. Let us suppose that considerations A, B, C, D, E, and F occur to her and that those are the only considerations that occur to her, and that on the basis of those, she decides to accept the job at Swarthmore. [...] Let us suppose though, that after sealing her fate with a phone call, consideration G occurs to her, and she says to herself: "If only G had occurred to me before, I would certainly have chosen the University of Chicago instead, but G didn't occur to me". Now it just might be the case that *exactly* which considerations occur to one in such circumstances is to some degree strictly undetermined. (Dennett 1978, pp. 293-294, *emphasis in original*)

The major feature of this model, according to Dennett, is this:

> When we are faced with an important decision, a consideration-generator whose output is to some degree undetermined produces a series of considerations, some of which may of course be immediately rejected as irrelevant by the agent (consciously or unconsciously). Those considerations that are selected by the agent as having a more than negligible bearing on the decision then figure in a reasoning process, and if the agent is in the main reasonable, those considerations ultimately serve as predictors and explicators of the agent's final decision. (Dennett 1978, p. 295)

Dennett's model of deliberative decision-making, accordingly, consists of two essential units: one is the consideration-generator whose functioning is sometimes undetermined; the other is the evaluating/selecting unit whose output is a decision or intention. The required indeterminism is embodied in the former, rather than in the latter. To appreciate the significance of the latter unit for our rational free agency, Dennett invites us to consider an analogy drawn on the following remarks of the poet, Paul Valéry:

> «It takes two to invent anything. The one makes up combinations; the other one chooses, recognizes what he wishes and what is important to him in the mass of the things which the former has imparted to him. What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it». (quoted in Dennett 1978, p. 293)

Consequently, the brand of libertarianism Dennett has recommended is also called "Valerian libertarianism" in the literature of free will (e.g., Bernstein 1989 and Double 1991).

1.2. Mele

In order to meet what libertarians want from indeterminism and to resolve some of the important problems they face, Mele suggests that

> it might be worth exploring the possibility of combining a compatibilist conception of the later parts of a process issuing in full-blown, deliberative, intentional action with an incompatibilist conception of the earlier parts. (Mele 1995, p. 212)

Mele proposes what he labels "modest indeterminism", according to which only some doxastic states or events are causally undetermined in deliberation:

> Some of [an agent's] beliefs will "come to mind", as we might say, and play a role in deliberation; other will not. But it is not causally determined which of these beliefs will come to mind and which will stay on the sidelines. Once a belief enters into the deliberative process, that "entering" event can play a role in the causal determination of subsequent mental events. Causally undetermined events can play a role in causally determining later events. (Mele 1995, p. 214)

What an agent judges best, we may suppose, is contingent upon which beliefs in a particular subset of his nonoccurrent beliefs "come to mind". And if the agent is an ideally self-controlled agent, then if he does make a decisive best judgment, he will consequently intend to act in accordance with the judgment.

The most notable feature of the model of deliberation based on doxastic indeterminism of the kind sketched, according to Mele, is that it does not diminish an agent's *proximal control* over her thoughts and actions, which is by stipulation compatible with the truth of determinism, and her responsibility as well, to any significant extent *in comparison with* attractive compatibilist accounts of free agency and moral responsibility based on determinism:

> [N]otice that we are not always in (proximal) control of which of our beliefs come to mind anyway, even if determinism is true. Assuming determinism, everything that happens on this front is causally determined, but the causal story often does not place the agent in the driver's seat. So, other things being equal, if responsibility for one's judgments is compatible with determinism, it is compatible, as well, with a modest indeterminism of the sort at issue. Plainly,

> which of our nonoccurrent beliefs come to mind can influence the outcome of our deliberation. An internal indeterminism that, as it happens, does not render us any less in proximal control of what occurs in this sphere than we are if determinism is true does not bring with it any direct impediment to responsibility for one's judgments that is not to be found on the assumption that our world is deterministic.
>
> This last point merits emphasis. One way to emphasize its significance is to make it a defining condition on the subset of one's beliefs that are subject at a time to indeterminism of the sort at issue that they are beliefs whose coming or not coming to mind is not something that one would control even if determinism were true. The agent who is subject to indeterminism in this sphere is not – simply on that account – worse off with respect to actual proximal control over his psychological and overt behavior than he would be at a deterministic world. (Mele 1995, pp. 215-216)

Moreover, this sort of internal indeterminism has the potential to supply what libertarians most want for free agency and moral responsibility. First, in being indeterministic, it seems to be sufficient to block the worry voiced in the consequence argument – the strongest argument for incompatibilism.[6]

> It allows for an agent's having more than one physically possible future and for its being true, on some incompatibilist readings of "could have done otherwise", that an agent could have judged, intended, and acted otherwise than as he did. (Mele 1995, p. 216)

Second, Mele suggests that

> the doxastic indeterminism at issue is an agent-*internal* indeterminism: it provides for an agent's having more than one physically possible future in a way that turns, essentially, on what goes on in him. (Mele, 1995, pp. 216-217, *emphasis in original*)

Mele goes on to argue that this sort of indeterminism can provide the grounds to secure the conditions for an agent's *ultimate control* over her choices and actions, which should not be fully subject to the causal determination of something external to the agent (e.g., the state of the world prior to the agent's birth together with the laws of nature). The notion of ultimate control, which is incompatible with determinism, preserves the crucial understanding that the origin or source of our free choices and actions is *in us* and not in anyone or anything else over which we have no control.

[6] See Van Inwagen 1983.

Both Dennett and Mele refuse to install indeterminism at other points of deliberation or at the exact moment of decision-making or even later. Dennett writes: «If there is to be a crucial undetermined nexus, it had better be prior to the final assessment of the considerations on the stage» (Dennett 1978, p. 295). For Mele, it is essential that except for certain doxastic mental states or events coming to mind indeterministically (which may emerge at any moment during deliberation), the rest of deliberation (the assessment of various courses of action and the formation of the best judgment), the formation or acquisition of the corresponding intention to act, and the agent's intentional acting accordingly, all proceed deterministically. And this is why the involvement of indeterminacy will not diminish an agent's

> proximal control over how he deliberates in light of the beliefs that do enter his deliberation. He may have considerable proximal control over how carefully he deliberates in light of these beliefs, over whether he deliberates in ways that violates his deliberative principles, and so on. (Mele 1995, p. 215).

According to the austere deliberative libertarian accounts advocated by Dennett and by Mele, the indeterminism required for an intelligible, plausible and coherent libertarianism works only in supplying input to an agent's deliberation.

2. Objections to Deliberative Libertarianism

This section surveys some major objections to deliberative libertarianism raised by a number of philosophers, including Randolph Clarke, Richard Double, Laura Waddell Ekstrom[7] and John Martin Fischer.[8]

2.1. The argument from luck

The problem of luck has become a main focus in recent debates about free will and moral responsibility, concerning both libertarian and compatibilist

[7] Ekstrom's view (2000) has been taken as a version of deliberative libertarianism (Clarke 2003, ch. 4). But as Mark Balaguer points out (2004, pp. 403-404), this interpretation is a mistake, which is partly due to Ekstrom's confusing and misleading terminology such as 'preference'. Balaguer reports that in private correspondence, Ekstrom has endorsed the non-Valerian libertarian interpretation of her view.

[8] Fischer (1995) presents and recommends a version of deliberative libertarianism as one that may satisfy libertarians in some respects.

accounts.[9] Ekstrom argues that deliberative libertarianism is not immune to this problem:

> [I]n my view, [Dennett and Mele] locate the indeterminism in the wrong place. Specifically, the views are too weak, in virtue of the indeterminism location, to secure agential freedom. On these views, free agents are subject to luck in what thoughts come into their minds as they are deliberating about what to do. But once the thoughts occur and the last of them has occurred during deliberation, there is a deterministic causal connection between the particular pattern of beliefs that has happened to occur and the subsequent decision outcome. But this is problematic. For I might be a free agent, on Dennett's or Mele's account, while being a victim, with regard to that I judge best and that I consequently intend and do, of what thoughts happen to occur to me at the time. Granted, there are "forks in the road" of some sort on this picture of free agency (alternate physically possible futures). But it is not up to me, the free agent, which one I take. Which one I take is decided by which considerations happen to come to mind, where this is indeterministically caused by some previous events. On both Dennett's and Mele's views, once a certain pattern of considerations has happened to occur to the agent, a particular action may follow of physical necessity and yet count as free. Since neither of the views includes an account of the nature of the self, they leave unanswered the question of why an act that is the causally necessary outcome of whatever considerations have happened to occur is plausibly claimed to be originated by the agent. (Ekstrom 2000, p. 137)

The objection from luck will be answered in section 3. Here I want to point out that Ekstrom's charge that both Dennett's and Mele's views lack "an account of the nature of the self" seems unfair. Even if the coming of certain considerations or beliefs to mind is a matter of mere happenings, over which an agent has no control, how the agent assesses, evaluates these considerations and reaches to the best judgment can nevertheless reflect the values, principles, preferences, capacities and habits that the agent possesses. Given the same pattern of considerations, different agents may well make different judgments and decisions, and the difference can hardly be accounted for without appealing to certain aspects of an agent's self. As Mele remarks,

> an agent's psychological condition (a combination of states and events) can be a central part of what causes his judging that it would be best to A, in a scenario in which the occurrence of a certain causally undetermined 'coming-to-mind'

[9] See Mele 2006 and Pereboom 2001.

> event just prior to the judging would have resulted in a different deliberative outcome. (Mele 1995, p. 217).

2.2. The argument from active difference-making

Clarke observes that a common belief about the freedom of will – one held by compatibilists and incompatibilists alike – is that in acting freely, agents make a difference to how things go by exercises of active control:

> The difference is made, on this common conception, *in the performance of a directly free action itself*, not in the occurrence of some event prior to the action, even if that prior event is an agent-involving occurrence causation of the action by which importantly connects the agent, as a person, to her action. On a libertarian understanding of this difference-making, some things that happen had a chance of not happening, and some things that do not happen had a chance of happening, and in performing directly free actions, agents make the difference. If an agent is, *in the very performance of a free action*, to make a difference in this libertarian way, then that action itself must not be causally determined by its immediate antecedents. In order to secure this libertarian variety of difference-making, an account must locate openness and freedom-level active control in the same event – the free action itself – rather than separate these two as do deliberative libertarian views. (Clarke 2003, p. 64, *emphasis in original*)

Deliberative libertarian accounts, Clarke argues, fail to supply this sort of difference-making. Dennett and Mele require that the coming to mind of certain beliefs, «which are not themselves actions», be undetermined, and allow «that these undetermined events, together with a nonactive reasoning process and its nonactive output (the making of an evaluative judgment), causally determine the decision» (Clarke 2003, p. 62).[10] On these views,

> [A]gent might be said to make a difference between what happens but might not have and what does not happen but might have, but such a difference is made *in the occurrence of something nonactive or unfree* prior to the action that is said to be free, not in the performance of the allegedly free action itself. Failure to secure *for directly free actions* this libertarian variety of difference-making constitutes a fundamental inadequacy of deliberative libertarian accounts of free action. (Clarke 2003, p. 64, *emphasis in original*)

[10] It seems better to use 'nonactional' to replace the term 'nonactive' in this quotation, for the latter may (wrongly) imply that the agent is passive in regard to her deliberation and decision-making.

2.3. THE ARGUMENT FROM DUAL RATIONALITY AND CONTROL

According to a general libertarian understanding of the condition of "could have done otherwise" or "alternative possibilities for action", when an agent acts freely, she must possess the capacity or power to act more than one way *deliberately* and *rationally*, rather than arbitrarily, capriciously, or irrationally, given exactly the same prior circumstances. This requirement is crucially different from and much stronger than what compatibilists usually demand – that the agent could have done otherwise *if* she had made another decision or choice. Whereas compatibilists interpret the power to do otherwise as a "one-way" hypothetical ability to choose otherwise than what the agent actually does, libertarians must impute to free agents a "two-way" or dual ability to choose otherwise, *in a categorical sense*. And for libertarians, this dual, categorical ability to choose or act otherwise must be exercised in a noncapricious and rational way (see Kane 1985 and 1996, ch. 7; Double 1991, ch. 1). Libertarians seem to

> be committed to the idea that free agents not only control which choices they actually make, but counterfactually *would* control alternative choices *had* they manifested their categorical ability to choose otherwise. (Double 1991, p. 15, *emphasis in original*)

In addition, as with dual control, when an agent makes a free choice, it should have been rational for her to have chosen another option under precisely the conditions that actually obtain.[11]

Double argues that deliberative libertarian accounts fail to "capture the spirit of the conditions of categorical ability to choose otherwise, dual control, or dual rationality, since it does not locate the indeterminacy where the libertarians want it, viz., at the final choice":

> To see this, compare Dennett's and standard compatibilist accounts. The latter hold that agents are free to decide otherwise, provided they would decide otherwise if they are so inclined. As we have already seen, the libertarians think that this hypothetical freedom is a sham. Now, Dennett's Valerian view holds that we do enjoy a categorical freedom to decide otherwise, since the appearance of some considerations on which we base our choices is literally indeterministic – that is, there are other physically possible worlds in which our

[11] Kane lately prefers the expression "plural rationality" to "dual rationality," and comes to see plural rationality as but one aspect of a more general "problem of plurality" for all libertarian accounts of freedom (Kane 1996, ch. 7).

> decisions would have been different. But this sort of categorical freedom, no less than the hypothetical freedom provided by the compatibilist's account, is too weak to satisfy the libertarian. [...] Libertarians want the freedom to decide either way, given the conditions that in fact obtain. So, although Dennett's view does an admirable job at producing one-way rationality – an unsurprising fact given that Dennett is a compatibilist – it fails to provide dual rationality, and it fails to produce the sort of indeterminacy that libertarians want. (Double 1991, pp. 200-201)

2.4. Indeterminacy and the problem of genuine control

A motivation for deliberative libertarianism is to solve the problem of agential control under the condition of indeterminism. Mele argues that the modest indeterminism he posits – internal, doxastic indeterminacy – is no worse than compatibilism in respect to proximal control, even if determinism is true. In addition, Mele suggests that installing indeterminacy in this way can preserve the crucial libertarian belief in alternative possibilities or freedom to choose and do otherwise. Fischer, in his insightful assessment of Mele's libertarian account, finds these claims puzzling:

> How can adding arbitrariness of the sort envisaged – the lack of determination of the beliefs that come to mind during deliberation – to a causally deterministic process yield genuine control? A libertarian of course will contend that an *entirely* deterministic process does not contain genuine control by the relevant agent. How, then, can installing the sort of indeterminacy envisaged – indeterminacy as to which belief states will come to the agent's mind – transform the sequence from one of lack of control to one containing control? This smacks of alchemy. [...] If an agent has genuine control in the sense of possessing alternative possibilities, he can make it the case that one path is followed, or another path is followed, *in accordance with that he judges best and chooses*. He can deliberately purse one course of action, or deliberately pursue another; what path the world takes (at least in certain respects) is "up to him". In contrast, when it is merely possible that something different have occurred, the path the world takes need not depend in the relevant way on the agent. In a genuinely random event, presumably there are various metaphysically open possibilities; but by definition no agent has control over what happens. (Fischer 1999, pp. 140-141, *emphasis in original*)

Fischer contends that, «whereas it may well be possible that Mele's libertarian agent do something different from what he actually does, it is not clear that he has genuine control over what he does». Given that the sequence of doxastic

states is not entirely determined by prior states of the agent, it is not clear that what the agent judges best and then does is genuinely up to him (Fischer 1999, p. 141).

Furthermore, Fischer argues that the deliberative libertarian account Mele advocates appears even worse than compatibilism in certain respects:

> [T]he compatibilist will point out that, even though the agent does not directly control what belief-states come to mind (in the sense of choosing them or willing them), they are envisaged as strongly connected to the agent's prior states to the extent that they are a *deterministic product* of those past states. Under determinism, one's prior states – desires, beliefs, values, general dispositions – *determine* the precise content and ordering of the subsequent doxastic states (that constitute deliberation), even if the agent does not directly control what doxastic states he will be in (and thus is not in the "driver's seat", in this sense). (Fischer 1999, p. 141, *emphasis in original*)

A similar objection is also raised by Clarke:

> It could be that, whenever one of us set out to make up her mind about which of several alternatives to pursue, all and only the most important and relevant considerations, or all and only those of this type that she had time to consider, would come promptly to mind, and these considerations would then figure rationally and efficiently in producing an evaluative judgment. In a deterministic world in which our deliberations always ran in this ideal fashion, we would exercise a valuable type of nonactive proximal control in deliberating. If chance at a later stage of deliberation would diminish proximal control, then chance of the sort required by Mele's view would seem to diminish this nonactive proximal control[...], anything that was found desirable in the independence secured by an account requiring chance here would have to be weighed against the loss of control in comparison with this deterministic ideal. (Clark 2003, pp. 68-69)

So far I have collected four major objections found in recent literature directed against deliberative libertarianism. [12] I shall reply to all these objections in the next section.

[12] There are some other worries against deliberative libertarianism. For instance, Kane suggests that selection from among chance-generated considerations «could not provide an account of moral or prudential choice», for «if responsibility is to be captured, then choosing morally or prudentially rather than from weakness of will could not merely be a matter of chance-generated alternatives» (Kane 1996, p. 162). Ishtiyaque Haji points out that whereas Mele's deliberative libertarian account «does make room for agent's having more than one physically possible future and for its being true that the agent could have judged, intended, and acted otherwise than she did», «such indeterminacy

3. Defending Deliberative Libertarianism

Before replying to the objections to deliberative libertarianism presented in section 2, I need to develop and refine the psychological model of deliberation and decision-making employed by this sort of indeterminism in several important aspects. The essence of the developments and refinements is to give

> agents a more active role in practical deliberation by way of efforts of will through which the agents might exercise greater control over the deliberative process – without eliminating the creative role of chance-selected considerations. (Kane 1996, p. 164).

3.1. Toward a realistic model of deliberation and decision-making

First, I suggest that the over-simplified, over-idealized indeterministic model of deliberation that has been implicitly assumed by most opponents of deliberative libertarianism should be abandoned. According to this simplistic model, the process of deliberation is in essence a linear, "one-shot" procedure: *after* all the considerations or beliefs, some of them are indeterministically caused, have come to mind as input to deliberation, all available alternatives are assessed and compared, and then a decisive best judgment falls out as the outcome of deliberation; *period*. Though this abstract model may be ideal for logical analysis of rational decision-making, it is a far cry from the reality of human psychology, leaving out many essential elements of an agent's efforts and control in deliberation. To see this, let us consider how a person is typically engaged in the process of deliberation. In the first round of deliberation, a set of considerations C_1 may come to the mind of the agent as input to deliberation; after all relevant options have been assessed and compared, a (tentative) best judgment B_1 falls out as a result. But the agent may deem that B_1 is unacceptable or unsatisfactory, or he may want to find an even better solution to the practical problem he is faced. He can readily embark on another round of the deliberative procedure: has another set of beliefs and considerations C_2 come to mind, and reach to another best judgment B_2 as a

does little to persuade us that the agent ensures that she has more than one physically possible future, etc.» (Haji 2001, p. 186). Since these worries have not been fully articulated, and I do not think they can amount to serious challenges to deliberative libertarianism, especially with regard to the refined psychological model of deliberation to be developed here, I will not attempt to silence them in this paper.

result. And the operation of this procedure can continue until a final decisive best judgment B is selected from among $\{B_1, B_2, ..., B_n\}$. As Mele remarks:

> The relevant indeterminism also applies, of course, to which nonoccurrent beliefs, in a certain subset of such beliefs, do or do not come to mind while deliberation is in progress. And even when an agent is on the verge of reaching a decisive better judgment, the (undetermined) coming to mind of a belief might prompt reservations that lead to reconsideration. So, in a scenario of the imagined kind, what an agent decisively judges best can be causally open as long as deliberation continues. Further, as long as deliberation is in progress it can be causally open when that deliberation will end, for it can be causally open whether a belief will come to mind and prolong deliberation. (Mele 1995, p. 217)

Or, as Robert Kane points out:

> Viewed in this way, ordinary practical reasoning or deliberation [...] is more like the trial-and-error processes of 'thought experimentation' that are characteristic of scientific discovery and creative problem-solving. The reasoner must consider various presuppositions and consequences of proposed lines of action, which usually involves the use of imagination to construct probable scenarios exemplifying those presuppositions and consequences. [...] As with instances of creative problem-solving, there are no fixed rules about what to consider, when one has considered enough consequences, and so on. (Kane 1996, p. 159)

A realistic human psychological model of deliberation is certainly much more dynamic, sophisticated and subtle than the abstract reasoning from $C=(C_1 \cup C_2 \cup ... \cup C_n)$ to B.[13]

Second, I think that the passivity of the coming to mind of certain considerations or beliefs in one's deliberation has been over-stated in the discussions of deliberative libertarianism. An agent is not always a helpless victim in regard to which subset of her nonoccurrent beliefs coming to her mind in deliberation. Consider Jones, the young philosopher in Dennett's example, who needs to make a choice between the positions offered by the University of Chicago and Swarthmore College. In her deliberation, it may occur to her that it is worthwhile to consult someone who has had first-hand

[13] Mele 1995, pp. 230-235 provides a nice case about the course of deliberation in which "intellectually sophisticated, self-reflective, self-assessing agents who seriously and responsibly tackle their decision problem".

personal experience with these institutions. Then she may perform a search in her memory in order to find out whom she may want to consult. Her recalling and searching for the particular items from her memory seem more like her (mental) actions, something that she actively, intentionally performs or brings about, rather than things that she merely undergoes or just happen to her.

In a recent article, Galen Strawson argues that in a fundamental respect, reason, thought and judgment neither are nor can be a matter of intentional action. «[M]ost of our thoughts – our thought-contents – *just happen*» (Strawson 2003, p. 228). But Strawson still allows an agent's mental acts to play a *prefatory*, *catalytic* role in thought:

> For what actually happens, when one wants to think about some issue or work something out? If the issue is a difficult one, then there may well be a distinct, and distinctive, phenomenon of setting one's mind at the problem, and this phenomenon, I think, may well be a matter of action. It may involve rapidly and silently imaging key words or sentences to oneself, rehearsing inferential transitions, refreshing images of a scene, and these acts of priming, which may be regularly repeated once things are under way, are likely to be fully fledged actions.
>
> What else is there, in the way of action? Well, sometimes one has to shepherd or dragoon one's wandering mind back to the previous thought-content in order for the train of thought to be restarted or continued, and this too may be a matter of action. We talk of concerted thought, and this concertion, which is again a catalytic matter, may be (but need not be) a matter of action: it may involve tremendous effort and focused concentration of will. Sometimes thoughts about the answer to a question come so fast that they have to be as it were stopped and piled and then taken up and gone through one by one; and this, again, can be a matter of action. Sometimes one has a clear sense that there is a relevant consideration that is not in play, although one doesn't know what it is. One initiates a kind of actively receptive blanking of the mind in order to give any missing elements a chance to arise. This too can be a matter of action, a curious weighted intentional holding open of the field of thought. (Strawson 2003, pp. 231-232)

Strawson's account of the prefatory, catalytic role of some mental acts in bringing certain thought-contents into mind makes good sense for deliberative indeterminism. An agent's performing of such mental acts of priming, attending, imaging and so on, which may well embody the agent's skills, habits and capacities of thinking and problem-solving, can make certain beliefs more or less likely to come to mind or consciousness in deliberation, though this

event is not entirely causally determined. This is quite in harmony with the spirit of Leibniz's familiar dictum "reasons may incline without necessitating": a person's skills and efforts in deliberation can positively, though not deterministically, influence the coming to mind of certain considerations.

Third, it is important to notice that once an agent is engaged in deliberation, it is up to the agent to decide when to terminate his deliberation, unless the process is interrupted from within or without. The purpose of deliberation is to find the best or a satisfactory solution to the practical problem that the agent is faced. But any deliberation is resource-consuming in terms of time, memory and cognitive capacity. In deliberation, a rational, resource-limited agent must consider whether to continue the deliberative process, that is, to have more beliefs and considerations come to mind and to make relevant assessments, in order to make more accurate assessments and find a better solution, or to terminate the process with the best available solution that has already found, in order to save the cost of deliberation. An experienced decision-maker would know that the temporal duration and mental effort devoted to deliberation do not guarantee the quality of decision-making. On the other hand, from the point of view of the deliberating agent, it seems sometimes quite uncertain whether or not that he has already selected the best solution for the practical problem at issue: perhaps just a little more effort, an all-round best solution will fall out. So an agent in deliberation may need to make hard choice under uncertain condition more often than usually conceived. It is thus up to him to decide when to terminate the deliberation, and thereby to make a practical decision on what to do. This mental event can be aptly viewed as a second-order decision: decide whether to terminate a deliberation. And this is something that a responsible agent must actively perform, rather than passively let happen to him.

I have attempted to improve and refine the psychological model of deliberation and decision-making in several aspects[14], which allows an agent to be engaged in the iterative processing of deliberation before making final decisions, to play an active role in bringing nonoccurrent beliefs into deliberative consideration, and to actively decide when to terminate a deliberation. We shall see below how these improvements enable us to response to the major objections directed to deliberative libertarianism.

[14] See Kane 1996, ch. 9 for suggestions and accounts of indeterminate efforts of will at other points in the deliberative process, which give agents a more active role in practical deliberation.

3.2. Replies to the objections

The argument from luck

I shall not attempt to tackle the vexed problem of moral luck[15], but only to show that the sort of indeterminacy introduced by deliberative libertarianism will not diminish an agent's control over his thoughts and decisions in comparison to that any qualified compatibilist account can offer. As noted earlier, deliberative libertarianism does not leave out "an account of the nature of the self" in an agent's practical deliberation and decision-making. First, the beliefs or considerations that come to an agent's mind in deliberation, including those caused indeterministically, are not generated from nowhere. They are what the agent has already collected and processed and still possesses. Second, the agent can make efforts, positively but not deterministically, to bring certain beliefs or considerations to come into deliberation. Third, the agent's assessments and evaluations of these considerations reflect the values, principles, preferences, and habits of the agent. Fourth, it is up to the agent to make the decision whether to terminate a deliberation with the best judgment already reached or to continue to search for a better option.

Nevertheless, despite his efforts, an agent may be still under the mercy of luck in regard to which beliefs coming to his mind. For instance, after the crucial beliefs and considerations coming to mind (indeterministically), Paul readily makes the best practical judgment and hence the best decision D; but Paul*, who is under the same prior conditions and shares with Paul the same set of values, preferences, and mental capacities, fails to reach the best decision D simply because the crucial beliefs and knowledge needed to reach the judgment have not come to his mind, in spite of his efforts. The difference between Paul's and Paul*'s decisions is solely due to their different luck. So indeed deliberative libertarianism is not entirely immune to the problem of luck. But as Mele has noted, «we are not always in (proximal) control of which of our beliefs come to mind anyway, even if determinism is true» (Mele 1995, p. 215). A psychologically plausible and realistic compatibilist account of human deliberation should not assume that in a deterministic world all relevant and important beliefs will consequently come to the agent's mind because everything entering into the agent's deliberation is deterministically caused. We can be forgetful about certain important information we already acquired,

[15] See Nelkin 2004 for a helpful review.

and we may even suffer from the frustrating phenomenon of tip-of-the-tongue, the feeling of knowing something that cannot be immediately recalled (see Brown 1991; Brown 2000 for reviews). Both deterministic and indeterministic account of human deliberation should leave room for such lucky events (for better or for worse) to occur. And there seems not point to assume that the sort of indeterminism introduced by deliberative libertarianism will render an agent worse off in terms of luck and control in this regard. So the problem of luck poses no special threat to deliberative libertarian accounts of free agency.

THE ARGUMENT FROM ACTIVE DIFFERENCE-MAKING

Clarke argues that in acting freely, agents can make a difference to how things go by exercising active control, «*in the performance of a directly free action itself*, not in the occurrence of some event prior to the action» (Clarke 2003, p. 64), and that deliberative libertarian accounts fail to supply this sort of difference-making. But as we have seen in the refined model of deliberation and decision-making developed above, an agent must decide when to terminate a deliberation, and this decision may well make a difference to how the agent will act consequently: if the agent decides to have more beliefs and considerations come to mind, in order to find a better alternative, she can readily do so, and this possibility is open to her. So the agent might be said to make a difference between what happens but might not have and what does not happen but might have, by directly exercising a mental act of deciding on whether to terminate her ongoing deliberation. Deciding is a mental act by nature, something that an agent actively performs rather than passively happens to her.[16] Therefore it follows that Clarke's attack against deliberative libertarianism in this regard is untenable.

Both Dennett and Mele insist that, in the model of deliberation adopted by deliberative libertarians, except for some considerations' coming-to-mind being caused indeterministically, all other stages of deliberation must be causally determined. So it seems obvious that, in accordance with their views, the very mental event of deciding to terminate a deliberation should also be deterministic. I would rather leave this question open, for it seems that a variety of libertarian views, including non-causal, agent-causal, and event-causal

[16] See McCall 1987; McCann 1998, ch. 8; Mele 2003, ch. 9 for arguments for the thesis that deciding is a mental action.

accounts other than deliberative libertarianism, can also make sense of this special second-order decision as a free mental act which terminates a deliberation.

The argument from dual-rationality and control

Richard Double argues that deliberative libertarianism does not qualify as an attractive libertarian account of free agency because it fails to «capture the spirit of the conditions of categorical ability to choose otherwise, dual control, or dual rationality» (Double 1991, p. 200), by which an agent can act more than one way deliberately and rationally, given exactly the same prior circumstances. And this is what a qualified libertarian account can offer whereas compatibilism cannot.

Double's charge, however, is largely misplaced. The alleged categorical ability to choose otherwise need not be exercised in every free action. According to one of the most compelling, intelligible and plausible libertarian accounts of free will which honor this sort of ability, namely, Kane's event-causal account, the exercise of this categorical ability usually implicates dual or plural conditions in terms of competing, conflicting or incommensurable motives, practical reasons, or values:

> Exercise of free will [...] typically involve incommensurable alternatives and incommensurable reason sets in one manner or another. In moral cases, the incommensurable reason sets are motives of duty versus self-interest; in prudential cases, desires for long-term goals versus present satisfactions; in cases of efforts sustaining purposes, desires to perform tasks or fulfill goals versus fears, inhibitions, aversions, and other countervailing inclinations. ... in practical deliberation also, agents are torn between competing and not easily comparable reasons for choosing between alternatives [...] The sets of reasons favoring each of the alternatives [...], the "incommensurable reason sets", comprise different and competing visions of what the agent wants to do or become. (Kane 1996, p. 167)

Notice that deliberative libertarian accounts have not incorporated plurality conditions into the psychological model of deliberation and decision-making: it has been presumed that all alternatives under deliberation can be accurately compared with each other and ranked accordingly. Whether this is a necessary simplification or unrealistic idealization, it would be question-begging to criticize deliberative libertarian accounts not being able to offer the categorical ability to choose otherwise typically exercised under the conditions of plurality.

Notwithstanding its failure to «capture the spirit of the conditions of categorical ability to choose otherwise, dual control, or dual rationality» (Double 1991, p. 200), deliberative libertarianism can still offer something that compatibilism cannot, and stand as an intelligible and plausible variant of libertarianism well worth wanting for its own right.

Indeterminacy and the problem of genuine control

Libertarians typically argue that in a deterministic world agents lack genuine control over their choices and actions. Fischer asks:

> How can adding arbitrariness of the sort envisaged [by deliberative libertarianism] – the lack of determination of the beliefs that come to mind during deliberation – to a causally deterministic process yield genuine control? [...] How, then, can installing the sort of indeterminacy envisaged – indeterminacy as to which belief states will come to the agent's mind – transform the sequence from one of lack of control to one containing control? (Fischer 1999, p. 140)

The reasoning that motivates Fischer's worry is this: since the envisaged agent lacks control over the events of (some) beliefs' indeterministic coming-to-mind during deliberation given that indeterminacy implies arbitrariness, deliberative libertarianism cannot do any better in securing genuine agential control than compatibilism. Indeed, according to the psychological model of deliberation and decision-making posited by deliberative libertarian accounts, the agent does not have the capacity to directly control *which* of her beliefs to be indeterministically prompted to come to her mind, but, as the refined model developed in this article has suggested, she can always decide and control whether to have *more* beliefs, some of them to be prompted indeterministically, come to her mind for deliberation. And this may have bearing on her final practical decision. Since it is up to the agent to decide when to terminate an ongoing deliberation, it is thus up to her and under her control whether to have more beliefs and considerations come to mind in order to envisage more alternatives and to make better assessments of the options. As some of the beliefs and considerations are indeterministically prompted, this sort of indeterminacy can thus constitute in the agent's certain kind of genuine control over her choices and actions which is precluded in a deterministic world.

Moreover, deliberative libertarianism helps to secure a sense of ultimacy that libertarians concern, namely the crucial understanding that the origin or source of our free, responsible choices and actions is in us and not in anyone or anything else over which we have no control. As noted earlier, Mele's notion of "ultimate control", by which an agent's performing a free action in the sense of ultimacy is not sufficiently caused solely by conditions external to the agent, is incompatible with determinism.

> Transformation of a deterministic actional process from one of lack of ultimate control to one containing such control by installation of the sort of internal indeterminacy that Mele recommends, should, consequently, not smack of alchemy. (Haji 2001, p. 183)

Fischer and Clarke both argue that the indeterminacy introduced by deliberative libertarianism seems to diminish an agent's control over her thought and deliberation in a certain way:

> Under determinism, one's prior states – desires, beliefs, values, general dispositions – determine the precise content and ordering of the subsequent doxastic states (that constitute deliberation), even if the agent does not directly control what doxastic states he will be in. (Fischer 1999, p. 141)

This helps to build a strong connection between one's prior psychological states and the deliberating process. And,

> in a deterministic world in which our deliberations always ran in this ideal fashion, we would exercise a valuable type of nonactive proximal control in deliberating. If chance at a later stage of deliberation would diminish proximal control, then chance of the sort required by Mele's view would seem to diminish this nonactive proximal control. (Clarke 2003, p. 69)

However, it is questionable whether it is *always* desirable and valuable for an agent to enjoy a strong, deterministic connection between her prior psychological states and "the precise content and ordering of the subsequent doxastic states" *in deliberating*. The purpose of deliberation is to figure out an optimal solution to the practical problem an agent is faced. This is sometimes a creative problem-solving task. The invoking of indeterminacy at certain points in this process may help to envisage new, novel ideas and alternatives that are not directly and strongly connected with one's prior psychological states. Furthermore, the whole process of deliberation is nevertheless under the agent's control: the agent can decide whether to allow more beliefs and

considerations come to mind and when to terminate a deliberation. So the working of indeterminacy is directed by the agent's purposeful executive guidance. And its effect can amount to practical decisions and actions only in accordance with the agent's overall psychological constitution.

Human creativity typically involves the generation of new ideas or concepts, or new associations between existing ideas or concepts, and results in producing or bringing about something novel, in imagining new possibilities not conceived before, and in seeing and doing things in a manner different from what was thought possible or normal previously. Creativity is not merely associated with the inspirations of geniuses in arts and sciences. It is also manifested in our ordinary daily lives, though in much less degrees of originality, ingenuity and significance. Thomas Edison once remarked that "to have a great idea, have a lot of them". The eminent chemist Linus Pauling echoed that "the way to get good ideas is to get lots of ideas and throw the bad ones away". It has been suggested that divergent thinking, which involves breaking away from what has been thought possible and normal, and flexible, novel generation of alternative solutions to a set problem, is a crucial element of creativity (Guilford 1967 and McCrae 1987). It is thus tempting to speculate that indeterminism may play a positive role in human creativity in general, and deliberative problem-solving in particular.[17]

Kane helpfully introduces the term "Taoist efforts" to characterize how agents "can willfully put themselves in a frame of mind that is *receptive* to new chance-selected considerations":

> Practical deliberators, like creative problem-solvers, do not have to wait for chance-selected considerations to occur in a manner that is completely uncontrolled and unbidden. When engaging in reflection about what to do, they can make efforts to relax their minds, freely associating and opening themselves to new thoughts and images that may well up from the unconscious. I call efforts of these kinds "Taoist efforts" because they are efforts to temporarily relinquish conscious control over thought process in order to be receptive to new considerations that may come to mind – that is, efforts-not-to-make-an-effort to control one's thoughts. Doors are thereby opened in deliberation that can free the mind from present commitments and ways of thinking. (Kane 1996, p. 165)

[17] Kane (1996, pp. 159-160) has offered some interesting and inspiring discussions on the relation between indeterminism, practical deliberation and creative problem-solving.

At the price of sometimes relinquishing total rational control of the conscious mind, as Kane suggests, there is room for indeterminism in the process of practical reasoning to enhance freedom and creativity: «This indeterminism make possible 'new beginnings' in practical deliberation that cannot be determined by reason, but can be used by it» (Kane 1996, p. 165).

4. Conclusion

In this paper, after developing and refining the psychological model of deliberation and decision-making employed by deliberative indeterminism in several crucial aspects, which allows an agent to be engaged in the iterative processing of deliberation before making a final decision, to exert some positive influence in bringing her nonoccurrent beliefs into deliberative consideration, and to actively decide when to terminate a deliberation, I have shown that the four major arguments directed against deliberative libertarianism are all untenable. Deliberative libertarianism survives these attacks as an intelligible, coherent and plausible libertarian account of free will that is worth being taken seriously.

Libertarians need to appeal to indeterminism to account for free agency. A principal challenge to this is that indeterminism, which implicates randomness, chanciness and arbitrariness, seems to undermine, rather than enhance, conditions for rational, responsible free actions. Deliberative libertarianism suggests a way to cope with this challenge: whereas an agent generally lacks control over *how* an indeterministic event happens, she can nevertheless control *when* to let a certain kind of indeterministic event to occur, *whether* to invoke more events of the sort, and *whether* to take into account the effects of these events. Moreover, as deliberative libertarianism has illustrated, indeterminism need not necessarily be "a *hindrance* or *obstacle* to our purposes that must be overcome by effort", as some libertarians grant (e.g., Kane 1999, p. 237, *emphasis in original*). Indeterminacy can nevertheless work as a positive and constructive ingredient that consists in human freedom, creativity and dignity.

REFERENCES

Balaguer, M. (2004). A Coherent, Naturalistic, and Plausible Formulation of Libertarian Free Will. *Noûs, 38*(3), 379-406.

Bernstein, M. (1989). Review of Robert Kane, Free Will and Values. *Noûs, 23*(4), 557-559.

Brown, A. (1991). A review of the tip of the tongue phenomenon. *Psychological Bulletin, 109*(2), 204-223.

Brown, S. R. (2000). Tip-of-the-tongue phenomena: An introductory phenomenological analysis. *Consciousness and Cognition, 9*(4), 516-537.

Buckareff, A. (2005). How (not) to think about mental action. *Philosophical Explorations, 8*(1), 83-89.

Clarke, R. (2002). Libertarian views: Critical survey of noncausal and event-causal accounts of free agency. In R. Kane (Ed.), *The Oxford Handbook of Free Will*, (pp. 356-385). Oxford: Oxford University Press.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

Dennett, D. (1978). On giving libertarians what they say they want. In D. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, (pp. 286-299). Cambridge, MA: MIT Press.

Double, R. (1991). *The Non-Reality of Free Will*. New York: Oxford University Press.

Ekstrom, L. W. (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.

Fischer, J. M. (1995). Libertarianism and avoidability: A reply to Widerker. *Faith and Philosophy, 12*(1), 119-125.

Fischer, J. M. (1999). Critical Notice of Alfred R. Mele's Autonomous Agents. *Noûs, 33*(1), 133-143.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York: McGraw-Hill.

Haji, I. (2001). Control conundrums: Modest libertarianism, responsibility, and explanation. *Pacific Philosophical Quarterly, 82*(2), 178-200.

Kane, R. (1985). *Free Will and Values*. Albany, NY: SUNY Press.

Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.

Kane, R. (1999). Responsibility, luck, and chance: Reflections on free will and indeterminism. *Journal of Philosophy, 96*(5), 217-240.

Kane, R. (2002). Introduction: The contours of contemporary free will debates. In R. Kane (Ed.), *The Oxford Handbook of Free Will*, (pp. 3-41). Oxford: Oxford University Press.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.

McCall, S. (1987). Decision. *Canadian Journal of Philosophy, 17*, 261-288.

McCann, H. (1998). *The Work of Agency*. Ithaca, NY: Cornell University Press.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology, 52*(6), 1258-1265.

Mele, A. (1995). *Autonomous Agents*. New York: Oxford University Press.

Mele, A. (2003). *Motivation and Agency*. Oxford: Oxford University Press.

Mele, A. (2006). *Free Will and Luck*. Oxford: Oxford University Press.

Nelkin, D. K. (2004). Moral luck. In *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/moral-luck/>

Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.

Strawson, G. (2003). Mental ballistic or the involuntariness of spontaneity. *Proceedings of the Aristotelian Society, 76*, 227-256.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

The Non-Mysterious Flesh: Embodied Intersubjectivity at Work

Liz Disley *
esd23@cam.ac.uk

ABSTRACT

Working one's fingers to the bone, having one's nose to the grindstone, *Knochenarbeit*... the metaphors we use for hard physical work are often applied equally to serious intellectual feats or exhaustive non-physical investigation or processing. In the phenomenological experience of work, what is the qualitative difference between physical and non-physical work? Hegel was the first to suggest a strong connection between work and sense of self-as-subject as among other selves, and his account in the master/slave dialectic and subsequent influential interpretations such as that of Kojève are focused on the physical process of 'negating' objects. Recent work on joint interests and joint attention focuses on goal-directed action that is paradigmatically non-physical, or where the physical aspect is incidental. In this paper, I investigate the role played by physical work in self-perception and in intersubjective relationships, specifically in a model of empathetic relationships. I also investigate the question of whether embodiment or shared goals and intentions are more important to a full account of intersubjectivity and empathy. As well as contributing to current debates about models of empathy, this discussion is also relevant to conceptions of solidarity and theories of the self in general, particularly as regards self-world relations.

Introduction – The Personal and the Professional

In this paper, I begin by examining briefly two accounts of more or less successful intersubjectivity and empathy in the work of Hegel and Husserl. The role played by the phenomenon of personal love between partners in these accounts has, in recent examinations, been seen as central to the extent to which they can be regarded as paradigms of successful and meaningful human

* University of Cambridge

interaction.[1] At the same time, the Hegelian and Husserlian accounts of human interaction which form the starting-point of my paper also have as an important element a framework of work as involving suppression of some type of desire and of community on the micro-level. My paper shifts the focus of the debate onto this phenomenon of work, asking whether group work, which I define more specifically in the next section, could in fact function as the arena wherein intersubjectivity and empathy can function ideally. My reasons for choosing to focus on work rather than love in this paper are threefold. Firstly, work is more easily examined on an everyday level as something which figures in the lives of all of us, certainly when the term "work" is broadly defined, as in my paper. Personal love is regarded, perhaps accurately, as a kind of mystical phenomenon which in many ways defies sober phenomenological analysis and is far distinct from typical concepts of reason. Secondly, personal love typically involves a very small community in each instance: paradigmatically, it involves a community of two.[2] Communities of workers, however, can be of any size. Focusing on work rather than love allows one to explore how intersubjectivity and empathy can and do function in wider communities. Thirdly, in work, it is much easier to isolate the physical aspect from the non-physical aspect and examine what role the nature of embodiment, and our understanding and experience of others' embodiment, can play in intersubjectivity and empathy. By "embodiment", I refer not only to the simple fact of a self being associated with a physical object, but the way the self experiences that body and other bodies.

[1] For examples of such examinations – not all of whom agree that Hegel or Husserl succeed in providing an account of successful human interaction – see e.g., Ormiston 2004, Williams 2000 and Hadreas 2007.

[2] I specifically use the term "personal love", rather than love in general, to exclude a variety of other phenomena we describe as "love" – love for art, cooking or a particular geographical landscape, love for my country or compatriots, and other types of love that cannot be described as having a particular object with which the one who loves has a direct and potentially reciprocal relationship. My love for the paintings produced by the Dutch Masters or Verdi's operas is not personal love in this sense because, as inanimate objects, they cannot respond to me. Equally, my love of an abstract concept such as a country – for it is surely some set of values, atmosphere or something else intangible that is the object of my love, not a tract of land between borders – can never be a relationship of personal love, since there exists no candidate for reciprocal action, leaving aside even the theoretical possibility of such action were such a candidate to be present. Love for compatriots is a case closer to the borderline, since I do at least want to leave open the possibility of altruistic love for one's fellow human being as a category of personal love. Nevertheless, love for compatriots in general is likely to have the same kind of very abstract character as love for one's country, and is not likely to focus on one specific individual who might reciprocate in a relationship of personal love.

After the discussion of Hegel and Husserl, I proceed with some definitions about the nature of work, and a more detailed discussion of how work might be the ideal arena for successful intersubjectivity and empathy. For reasons of space and scope, it is not possible to provide here a full examination and justification of a particular definition and understanding of what might count as "successful" empathy and intersubjectivity, so I will offer just some brief thoughts. First of all, to define the terms themselves, I define intersubjectivity in the following way; as that quality of the external world and/or human minds that allows us to see other minds and their attendant bodies and subjects and objects in the fact of our own subjectivity and objectivity, as willing, perceiving and acting subjects like ourselves inhabiting the same world.[3] Empathy is, as the original German term "*Einfühlung*" (feeling-in or -into) suggests, more closely connected with our ability to access and understand the motivational, emotional and affective states of those who also inhabit our world. Degrees of "success", in the first instance, are therefore concerned with the degree and complexity of understanding of these types that is achieved. On this analysis, neither intersubjectivity nor empathy is a binary quality in the sense of either being achieved or not being achieved, but allows for a wide range of degrees. There is certainly also an ethical dimension concerned with what obligations there might be on us to achieve successful intersubjectivity and empathy, or whether, to argue from a different angle, the fact that intersubjectivity and empathy are possible entails certain ethical demands. These extremely useful and pertinent questions are not my concern here, but would certainly form an interesting basis for a discussion.

1. Work, Desire and Purpose – Some Definitions

What is the central and essential phenomenological quality of work? Certainly, we use the same kinds of metaphors involving physicality (working one's fingers to the bone, *Knochenarbeit* in German) or even animals (working like a dog, donkey-work) to describe this experience or the observation of someone else's work. However, the terms we use to describe our own work or, generally approvingly, the effort of other people, do not mean that the experience we have of our own or of others' work does not vary greatly depending on whether

[3] Note that this general analysis is agnostic about whether intersubjectivity and empathy are made possible by some biological or even ontological feature of the world or human subjects.

this work entails physical effort. Equally, whilst the colloquial terms above paradigmatically conjure up the image of lone, heroic toil, this does not necessarily contribute anything to the question of whether there is a phenomenological difference in undertaking work alone or in a group and, if so, how this difference can be characterised. The first part of this paper will suggest some differences in terms of phenomenological experience of physical and non-physical work, both undertaken alone and as part of a group. In this analysis, my aim is to discover what, if any, part is played in the phenomenological experience of sole or group work by embodiment.

Any analysis of the phenomenon of work needs to volunteer at least a provisional definition of the boundaries of that phenomenon. Certainly, it would be unwarranted to assume from the outset that the definition of "work" which will be important for this particular analysis is co-extensive with the definition of "work" from a socio-economic point of view. Indeed, the provisional definition with which I am working is broader than that. The economic definition of work as a sustained task undertaken in order to earn money will not be sufficient for this analysis, or at the very least the financial benefits of work will not be seen as the most phenomenologically pertinent feature of the phenomenon.

Many critics of, and commentators on, Hegel and the subsequent idea of work, alienation and freedom have emphasized one particular feature, that is, the extent to which work is a suppression of one's natural desires. A fairly neat analysis can be made with the help of Harry Frankfurt's first- and second-order desires, where "natural desire" is understood to be roughly equivalent to first-order desires.[4] A first-order or natural desire is a straightforward, immediate desire, formed with minimal, if any, cognitive involvement – the desire for rest, shelter, food or drink. A second-order desire, on the other hand, is a desire for a desire and thus not a simple natural desire. For example, one might desire fame or fortune, but this is likely to require the suppression of natural desires such as that for rest. One might desire, then, that one does not desire rest. That is to say, the first-order desire is held in check in order that the second-order desire can take precedence. Work involves, but is of course not limited to, holding one's natural desires in check in favor of some purpose that extends beyond these natural desires.

[4] Cf. Frankfurt 1988.

1.1. Basic and higher purposes

This desire criterion concerns partly what work is *not*, in terms of purpose. It is at the same time important to point out that work is for some purpose, and must be understood by the worker as being for some purpose. This can be understood as a two-fold requirement. First of all, a particular task of work should be understood as purposive by the worker – that is, if I am standing at the front of a room making some comment about Kant, I should understand that the purpose of this task is to teach my students about Kant, perhaps so that they can pass their exams. Equally, in the case of physical work, I should understand that my polishing of a window is for the purpose of cleaning that window, perhaps to improve the appearance of a building. The work should have what I will call basic purposiveness (the basic purposiveness criterion). I will examine whether, for joint or collaborative work, it is important for work that is relevant from the point of view of my analysis of empathy and intersubjectivity that each participant has the same broad idea of the basic purposiveness of the work. As well as basic purposivenss, there is the question of higher purposiveness (the higher purposiveness criterion). This is rather more difficult to define, and it will be a major task of my paper to examine what this consists in and to what extent it is important to the phenomenon of work as key to an understanding of empathy and intersubjectivity. Higher purposiveness as I am defining it concerns the perceived overall purpose of the task as it contributes to the person's job or profession, or wider significance of a task that does not fit into the framework of employment, for example, a small task that is part of the wider purpose e.g., of renovating one's home.

In the simplest possible terms, I understand a "worker" as being someone engaging in a task which involves action other than as a function of one's natural desires (the desire criterion) which also involves some kind of basic purpose (the basic purposiveness criterion). For individual work, I initially leave open the possibility or necessity of a higher purpose to the task at hand. This is a fairly minimal description of work, which would certainly extend to running errands or performing basic household tasks, as well as, importantly, caring for another person or performing fairly simple acts of kindness (holding open a door, for example). For group work, I initially work with the following definition – the definition of individual work, but with another individual worker that assists one in fulfilling either a basic purpose (e.g., preparing a meal for one's family) or a higher one (providing the family with a nutritious and pleasant-tasting diet). Additionally, for the purposes of the

phenomenological analysis of joint experience, I will stipulate for joint or group work that there must be some kind of sustained communication between workers in a group work situation, leaving aside for the moment the question of whether the workers have to have physically met at some point.[5]

2. Hegel and Husserl – Work and Love

Before I begin the analysis of physical and non-physical work using the phenomenological method, I will briefly sketch out the historical background to this question. Perhaps the most famous example of a phenomenological analysis of physical work is provided by Kojève's reading of Hegel's master/slave dialectic in paragraphs 178-196 of the *Phenomenology of Spirit* in his *Introduction to the Reading of Hegel* (Kojève 1969) Kojève offers us an analysis of work as a process that limits the slave's freedom still further by holding him in fear of death, but in fact proves his salvation as the interaction with the external world distinguishes him from animals and humanizes him. For this type of work, as I will explain, the physical aspect of work is central to Kojève's narrative of liberation which can in turn be casts as a comment on the objectivity and subjectivity of the human self. At the same time, if we are to take the core of Kojève's analysis seriously, what he says about the liberating and humanizing power of work is even more convincing if we are considering physical work in a group rather than lone work. The other central historical figure important for the purposes of this paper is Edmund Husserl. Husserl, in his *On the Phenomenology of Intersubjectivity* (Husserl 1973), develops an account of communal striving (*streben*) that forms the basis of personal love, but which can potentially be widened out to form a kind of ideal model of positive ethical intersubjectivity. However, although embodiment is certainly a concern for Husserl in the *Cartesian Meditations* (Husserl 1950), the *Ideen* II (Husserl 1991), the Krisis writings (Husserl 1976) and Husserl 1973, his account of intersubjectivity and empathy does not fully account for the special role played by one's own experience of one's own body as an *object* in the lifeworld as well as a subject that is the geographical centre (the "absolute

[5] A full examination of the phenomenological experience of modern workplaces with electronic communication, video-conferencing and the like as opposed to traditional workplaces where colleagues are physically together for the majority of the time is unfortunately beyond the scope of this paper, but this would certainly be a worthwhile candidate for a fuller study.

here" of e.g., Husserl 1950, p. 146) of self-government.[6] To explain what role embodiment needs to play in the analysis of work, I refer later to the work of Maurice Merleau-Ponty.

Hegel and Husserl provide us with two accounts of working together, understood in the broadest possible terms. Their two accounts will therefore form the starting-point of my enquiry. Choosing these two accounts in particular requires an engagement with the general question of intersubjectivity and monistic ontology to which I shall return at various points in the paper. Both Hegel and, less directly, Husserl, are accused of subsuming true intersubjectivity into the monistic unity of subject and object. The criticism of the former is lead by Michael Theunissen, and of the latter, Max Scheler.[7] Both Hegel and Husserl's accounts of intersubjectivity appear to treat the unity of marriage partners or lovers as a kind of paradigmatic example of effective realisation of that phenomenon. Despite the lack of obvious similarity between work and love, in fact these examples both depict a unity with some kind of shared intentions that is grounded on a deeper unity of consciousness, which may or may not imply or require a deeper ontological union.

For Hegel, married couples form the smallest unit in a civil society shaped by an intersubjectivity of reason and action. Some recent scholars have suggested that it is in a loving relationship that Hegelian intersubjectivity has its most positive and well-functioning expression as perfect mutual recognition (in the technical sense of *Anerkennung*) of the other partner's ontological status. Briefly stated, recognition in Hegelian terms is specifically the recognition of the Other as having a particular ontological status that is the same as one's own, and, crucially for Hegel's account, it must be mutual. One cannot recognize without being recognized, and vice versa. Recognition, for Hegel, is a necessary part of the development of self-consciousness.[8] Whilst

[6] The prolific nature of Husserl as a writer as well as the specific historical challenges of tracing his shifting views render many accounts of a particular concept or view of his open to challenge from an earlier or later work. Whilst I do not claim that Husserl has one constant view of intersubjectivity, I do assert that the elements of his concept that are particularly relevant to the concerns of this paper remain sufficiently constant for this not to constitute a serious objection from other works.

[7] See Theunissen 1991 in general and Scheler 1970, p. 75.

[8] This is not an uncontroversial account of Hegelian recognition or the development of self-consciousness. John McDowell has recently followed Joseph Flay (Flay 1984, p. 86) and George Armstrong Kelly (Kelly 1984) in advancing a view of the master/slave dialectic as an internal process

the master/slave dialectic describes a failure of mutuality and therefore a failure of recognition, the loving relationship describes a relationship where the two partners live in recognition and harmony. Others have suggested that the ascription of a unity of consciousness to the loving couple, as well as Hegel's monistic ontology in general, involves a subsuming of genuine intersubjectivity into one monistic substance, thereby rendering the idea of a social construction of reality incoherent or impossible. Indeed, the famous quotation from the Philosophy of Right initially seems to lend some weight to the assertion that individual consciousness is subsumed:

Love means in general terms the consciousness of my unity with another, so that I am not in selfish isolation but win my self-consciousness only as the renunciation of my independence and through knowing myself as the unity of myself with another and of the other with me.[9]

Both in German Idealism and in the work of Husserl, intersubjectivity is not simply, as Allen Wood describes it, «our conception of the mentality of others and our awareness of it», but accords much more closely to the definition given in the first part of this paper (Wood 2006, p. 66). Hegel's and Husserl's accounts of recognition and empathy both clearly require that the subject is in some way an actor in a community, and it is precisely in this sense that it can be useful to use close relationships between two people as a model for human interaction in general. What does seem clear from the *Philosophy of Right* is that recognition is a process with three clear steps, and not a case of simple desire for mastery or subjectivity as in, for example, a Sartrean account. In fact, the first desire is that for objectivity, to disappear into the other person, which is then replaced by a desire for subjectivity and then an achievement of both objectivity and subjectivity in the eyes of oneself and the other. The lover is not subsumed, but recognized, and the simple desire becomes a complex one. The simple desires for objectivity and then subjectivity must be held in check and suppressed in order for the more complex desire for recognition to emerge. This notion of desire held in check is one that can be observed on the simple empirical level of any close relationship between two individuals where each individual wants something for the other individual as well as for herself.

The main difficulty with Hegel's account of recognition and intersubjectivity is the centrality of the master/slave dialectic which is open to

involving not a distinct other, but rather the finding of oneself in one's formative activity and the move from theoretical cognition of life to a practical immersion in it. See McDowell 2009.

[9] Hegel 1991, addition to paragraph 158.

such a wide range of empirical elucidations. Husserl's account of an intersubjectivity of action is far more empirically comprehensible. I will turn now to the issue of the role played by love in Husserl's account of intersubjectivity. Husserl's account of personal love proceeds from his general analysis of the concept of communal striving.[10] Whilst love is more than communal striving, this working together forms the basis of close interpersonal relationships. His general concept of empathy forms the basis for his discussions of communal striving and personal love. The key concept is *Nachverstehen* or, as Peter Hadreas translates it, understanding-following-after-another (Hadreas 2007, p. 20). *Nachverstehen* is a kind of empathetic understanding of one person following after another which makes the other person impossible to objectify. Part of the refusal to objectify the beloved is due to the appreciation of her particular subjectivity – an object could simply be replaced.

There are two particular features of this phenomenon that demonstrate the clear connection to Husserl's broader intersubjectivity and an account of ethical love in the community. Firstly, there is this impossibility of objectifying the beloved; as Peter Hadreas puts it, «The beloved person remains more than can be collated into an object» (Hadreas 2007, p. 20). This has clear parallels with the irreducible nature of the community. Secondly, there is the emphasis on communal striving and activity in the couple as well as in the wider community, as the person of the beloved is disclosed to the other part of the couple through sharing in his acts and following in his footsteps, either cognitively or literally. Working together for common goals is crucial:

> As one who loves I know that, whatever I think, feel, strive for, or do, all are necessarily 'in the interests' of my beloved, is right for the beloved, and is right for the beloved not only in the sense of my not being scolded by the beloved, but rather as something I strove for in the interests of my beloved's striving. (Husserl 1973, p. 173)[11]

The higher purposes of the beloved and the lover are completely at one with each other – all goals, whether or not they are basic or higher purposes, are at one, because the lover is in love with the unique beloved. The question remains whether this complete meshing of goals of action could persist with a weaker bond, for example, that of the workers. Leaving aside the question of whether

[10] See e.g., Husserl 1973, p. 171.
[11] Adapted from a translation in Hadreas 2007, p. 37.

Husserl's account is simply too demanding, I will move on to examine an account that goes beyond Husserl's concept of the self and the body and places embodied self at the very centre of intersubjectivity and empathy.

3. The Worker as Embodied – Merleau-Ponty and Skill

What is the special relevance of embodiment to work? How does the fact that we are, as Merleau-Ponty puts it, "psycho-physical subjects" affect us as workers, and how does this relate in particular to work undertaken in a group? The following comment is key:

> In so far as I have hands, feet, a body, I sustain around me intentions *which are not dependent upon my decisions and which affect my surroundings in a way which I do not choose*. These intentions are general [...] they originate from other than myself, and I am not surprised to find them in all psycho-physical subjects organized as I am. (Merleau-Ponty 1962, p. 440, *my emphasis*)

Not only does the simple fact of my embodiment mean that I am not able to predict with certainty how my goal-directed action will translate into the desired result on even the most basic level, but due to my physicality, my *intentions* do not depend on my decisions because even those intentions have to be developed with regard to the physical environment of which I, *qua* physical being, am part. Merleau-Ponty's comment also touches on the concept of intersubjectivity of embodied objects, albeit in a fairly minimalistic sense – there is a simple reasoning that because I am limited in translating my decisions into concrete intentions (and therefore am not radically free in the sense that Sartre would insist I am), then others whom I identify, for whatever reason, as being crucially similar to myself, must also be limited by such circumstances.

Seen in the light of the current discussions of models of empathy, Merleau-Ponty's comments about intentions, decisions and the understanding that other psycho-physical objects are similarly limited in their goal-directed activities are particularly interesting, especially given his concept of skilful action. According to Merleau-Ponty in the *Phenomenology of Perception*, there is a strong connection between the physical self and the world (the intentional arc) which means that when the active body acquires new skills, these are stored not as mental representations but as dispositions which allow one to respond to one's physical environment – what we would, in the common

vernacular, call "skills". Moreover, the "maximal grip" is the process which allows the active body to refine its skills and bring the physical situation closer to what it regards as the optimum – Merleau-Ponty uses the example of a painting which has an optimal distance from which it should be seen. The vital point about skill in the intentional arc and maximum grip is that the capacities developed there are not propositional knowledge. Can we equally have skills with regard to other psycho-physical objects, and, if so, do these capacities treat such psycho-physical objects simply as part of the external furniture, or do they respond at the same time to the non-physical aspects of the Other?[12]

The significance of this question becomes clear when one considers the modern debate in the theory of empathy between theory-theorists and simulation theorists. Broadly speaking, the two views can be summarized in the following manner. The theory-theorist sees empathy as involving a theory of mind that is held by the empathiser which allows them to attribute intentional states to the person with whom they are empathising. In other words, for the theory-theorist, empathy involves propositional knowledge, unlike Merleau-Ponty's account of the intentional arc and of maximum grip. The simulation theorist, on the other hand, believes that by using our cognitive capacities, we put ourselves in the position of the other and simulate their mental states in ourselves. As one of the early proponents of this theory, Jane Heal, puts it, «we take the subject matter of that thought, whether we believe the same or not, and think directly about it» (Heal 1995, p. 35). In this sense, simulation theory does not involve propositional knowledge – in fact, the factual contents of our beliefs are, as she points out, irrelevant from the point of view of our empathising. Certainly at first glance, it seems that empathy, for the simulation theorist, is a kind of skill, even if it does not follow the precise path of skill-development traced out by Merleau-Ponty in the *Phenomenology of Perception*.

If it is the case that empathizing in general is a simulation process not involving propositional knowledge or a theory of mind, then one could argue that the skill of working with another person, two workers both limited by the fact of their being physical objects amongst physical objects, is rather like the example of a chess player developed by Hubert Dreyfus in a 2002 paper. Dreyfus describes the stages that a chess player learning to play to a very high

[12] I follow a general convention used by a large number of writers on intersubjectivity of capitalising the word Other when it refers specifically to a candidate for intersubjective relationships.

standard goes through – first the simple memorisation of playing rules and possible move permutations, moving on to a stage where those rules work together with each other in a kind of interplay with a developing skill that does not require conscious reference to propositional knowledge, finally ending up at the following stage where expertise is used all the time and the body of knowledge is referred to only occasionally. This allows the immediate intuitive situational response that is characteristic of expertise (Dreyfus 2002, p. 372). If this analysis is applied to work in a framework influenced by Merleau-Ponty, we begin to get a general picture of skilled working with others that does not primarily require reference to a body of propositional knowledge. Whatever the degree to which the chess analysis, to which I return in the fourth section of this paper, can be applied to the world of work, the account of skill in general draws us closer to the conclusion, as Wringe 2003 puts it in the context of a simulation account of empathy, that «[o]ur beliefs do not constitute the sum or even, necessarily the most important part of our mental lives» (Wringe 2003, p. 354). Empathy is rather more than the attribution to others of mental states.

3.1. Work and intercorporeity

In the previous subsection, I suggested that Merleau-Ponty's account of skill presents a picture of the phenomenological experience of individual and group work that fits well with the simulation-theory view of empathetic relations. In this subsection, I will briefly examine Merleau-Ponty's own view of work and recognition in the master/slave dialectic in the *Phenomenology*. David Storey argues that, in terms of arguing for more fundamental structures of consciousness as an explanation of human experience, Merleau-Ponty is to Husserl what Hegel is to Kant (Storey 2009, p. 62). According to Storey, both Merleau-Ponty and Hegel are fundamentally concerned with restoring the great chain of being by re-imbuing what are often seen as non-philosophical objects with ontological significance and making clear the fundamental unity of self and object in a pre-conscious sense. In Hegel's case, of course, that is motivated by a commitment to monistic ontology in general. This of course brings forward the well-established question of whether Hegel's objections to Kant (and, by analogy, Merleau-Ponty's objections to Husserl) take as their primary ground the fact that Kant/Husserl's general framework provides an empirically or phenomenologically insufficient account of human experience, or the fact that the ontological presuppositions are faulty to begin with. Storey suggests that the difference between the two pairs of philosophers relates to

their general attitudes towards monistic and dualistic ontologies. One proposition, which I can explore only briefly in this paper, is whether effective intersubjectivity (or a fully descriptive account of intersubjectivity) in fact requires some kind of monistic ontology. In other words, it could be that the overcoming of the dualism which Storey characterizes as that of Spirit and Flesh will in fact require an overcoming of other dualisms, most fundamentally of all that of subject and object. I will put this question aside for the moment and return to it later.

The key concepts which differentiate Merleau-Ponty from Husserl are those of intercorporeity and skill. Whilst Husserl pre-figures Merleau-Ponty in terms of his concern with lived, bodily experience, his account of embodiment, certainly in the *Cartesian Meditations*, focuses on the Other's governing of one's own body as similar to my governing by own body rather than the body as a limitation on freedom which must constantly adapt to obstacles to performing a particular desired action as in Merleau-Ponty's account of skill.[13] Husserl's account of spatial subjectivity in Husserl 1973 in particular is full and detailed, but it demonstrates an important limitation which is crucial for the account of intersubjectivity and work, namely the inability to fully "objectivate", as Peter Reynaert puts it, my body as a whole (Reynaert 2001, e.g., p. 211). In order to experience my body as an object, I would have to step outside that body and assume different perspectives from the one I occupy.[14] This has important consequences for the self at work and for access to experience of the Other, which in Husserl's account of embodiment and Paarung seems rather theoretical in terms of comparing data.[15] Not only is it impossible for Husserl to develop an account of something like Merleau-Ponty's skill, he also cannot develop even a basic account of something like the later philosopher's concept of crisscrossing described below, as, for Husserl, we can objectivate parts of our bodies but not our bodies as a whole. Therefore, the methodological explanation breaks down.

Merleau-Ponty's concept of crisscrossing, where we experience embodied others and ourselves as both objects and subjects by shifting focus between our left hand touching our right hand and our right hand being touched by our left hand, is a prior stage of embodied intersubjectivity before that of full

[13] See e.g., Husserl 1950, p. 128.
[14] Cf. Husserl 1973, p. 413.
[15] Cf. Husserl 1950, p. 147.

intercorporeity.[16] As mentioned, it is always going to be a further question *why* this should be the case. What is special about the other person's right hand that I should be able to apprehend it in a similar way to my own hand? Why don't I encounter it in the same way as a door handle or a hockey stick? If it is simply a matter of physical similarity, then intersubjectivity on many definitions has not been achieved at all, and certainly I am thinking about and empathizing with the Other very much on the level of the theory-theorist, basing my conclusions on a theory of mind and the ascription on the basis of physical similarity to the Other of propositional knowledge about her cognitive faculties. For a number of reasons mentioned above, this is deficient in terms of empathy, if not also in terms of intersubjectivity, and therefore deficient on both the phenomenological and the ontological levels.

Merleau-Ponty's concept of intercorporeity provides us with an account of intersubjective embodiment that attempts to explain without the use of theories of mind or propositional knowledge why it is that, as succintly puts it, «the flesh of another person is not an absolute mystery» (Brubaker 2000, p. 96). Merleau-Ponty's account of incorporeity is one which has experience, and not propositional knowledge, at its heart. He gives the example of the left and right hands as compared with the hand of another, and poses the following question: why «when touching the hand of another, would I not touch in it the same power to espouse the things that I have touched in my own?» (Merleau-Ponty 1964, p. 141). Merleau-Ponty links this to color perception or apprehension. When I think of my own experience of the color green, I recognise that this is somehow a private experience not transparent to the Other. At the same time, however, I recognise it on reflection as a pre-cognitive apprehension and not a judgement in the Kantian sense. As Brubaker puts it,

> by witnessing the sensuous flesh constitutive of our own *idios cosmos,* each of us may posit, by analogy, "another presumptive domain of the visible and the tangible" that cannot be expressed in the languages of physical bodies and intentional consciousness. (Brubaker, 2000, p. 96)

We come to this conclusion, or, to put it more accurately, we experience the Other in this way, because of the way we experience our own body as a perceiving body and because we can ascribe to another experience we recognise as being private.

[16] See e.g., Merleau-Ponty 1964, p. 135.

If we accept Merleau-Ponty's account of intercorporeity, what kind of consequences does this have for the intersubjective experience of work? Does it demonstrate that intersubjectivity and empathy will somehow work better in physical group work than in non-physical group work? One thing to notice in particular about Merleau-Ponty's account of intersubjective embodiedness is that it does not primarily involve goal-directed action, but rather simple encounters with the physical Other. At the same time, however, if we also take into account Merleau-Ponty's comment that a crucial part of embodied is the experience of being frustrated in some way by one's physical environment purely in the sense that one is not, *pace* Sartre, radically free to realise one's intentions, we can begin to see how intercorporeity might be transferred to the work realm. What is crucial for incorporeity is some kind of realm, a cosmos in which one acts. A work environment is a specific example of such a cosmos. I witness that my physical experience of the workplace is private and opaque in some important way for the Other, but that they have a similar cosmos which they experience in some broadly similar way. Their physicality is not a complete mystery for me. How can I make it even less of a mystery? Presumably by physically standing in their cosmos and sharing physical experiences. Two workers performing similar physical tasks side-by-side will get as close as anyone can to each other's physical experience, but the importance of shared goals, crucial for Husserl, pales into insignificance on the Merleau-Pontian account. Indeed, the experience of incorporeity would be equally strong in the case of two exact competitors, for example, two runners competing would have a stronger bond of this nature than competitors in a relay team. This is not necessarily a deficiency in Merleau-Ponty's account – indeed, a strong moral dimension has been observed in his account which I shall discuss later in this paper. In many ways, it is Merleau-Ponty's account which might be seen as the one that would most easily account for the phenomenon of solidarity.

It is a further question, of course, whether this incorporeity points towards or even requires some kind of general ontological framework which enables us to extrapolate from our experience of our own physicality to that of others, and I will return once more to this question later in the paper. I will note here that the notion of intercorporeity seems to strongly support the simulation theory of empathy as expounded by Heal *et al.* whilst seeing only a minor role for goals and intentions. It is to the notion of shared intentions that I will now turn.

4. Shared Intentions

One good candidate for a view which opposes that of Merleau-Ponty in almost every respect is John Searle in his account of shared or collective intentionality. Searle's central claim relevant to this paper in his *The Construction of Social Reality* is that collective intentionality is not at all dependent on even the existence of a world outside the mind.[17] His argument goes broadly like this: collective intentions exist only in individual brains, but it is nevertheless possible for individuals to have a so-called we-intention because of a kind of "shared Background", capitalized because it is being used in a technical sense to mean conditions necessary for certain cognitive activities and, crucially, language. Whilst Searle's account might seem radically individualistic, in fact he argues that the having of a Background sense of relevantly similar others is, in fact, inborn and something we have in common with biologically similar species (Searle 1995, p. 414). This provides an interesting counterpart to the idea of some form of ontological unity, namely a kind of biological unity, or at the very least some kind of biologically-determined access to the Other at least in terms of their cognitive faculties. Whilst this would be likely to fall short in terms of empathy in as far as empathy involves some kind of shared emotion, it seems to be a good candidate for practical intersubjectivity. Certainly, to put it in Heideggerian or Sartrean terms, it is a form of pre-reflexive consciousness – there can be no reflection when recognizing the kind of cognitive capacities the Other has based on some biological consciousness. When making this particular point in a paper about the intersubjectivity of meaning, Carlos Cornejo makes the following point:

> In natural circumstances I am not in front of others as they were objects being-present-at-hand. Instead, we usually are actively engaged with them in common activities, so that their behaviors seem us pristine and fullfledged of meaning. Within the minimal communicative situation, the other is from the start available, not present-at-hand (Cornejo 2008, p. 175).

Whilst it is not the fact that I am engaged with the Other(s) in some common activity that allows me to draw conclusions about their Background, the intersubjectivity of meaning is something that is meaningless without common activity. The most obvious illustration of this point is Searle's own example f money – money is only money (that is, only has monetary value)

[17] See e.g., Searle 1995.

because of some tacit agreement that we will all accord it this meaning and accept it as such. So, whilst it is possible to have a we-intention as an individual even as a brain in a vat if that individual posits the existence of others who hold this intention, it is only when genuinely engaged in common activities that intersubjective meanings can actually come into existence. Meijers and others enforce this point with the objection that Searle does not take into account the extent to which collective intentions are rule-governed (Meijers 2003). Joint action is essentially normative. If we examine again Searle's own example of the football team who have some joint intention or joint goal, the forming of the intention to play football and perhaps beat the other team involves the forming and accepting of some kinds of rights and obligations which all of the players, at least in broad terms, understand. It is not in the least bit meaningful to speak of these norms if the person who has formed the we-intention is a brain in a vat. Whilst the kind of we-intention Searle describes might be sufficient for collective intentionality in the narrow sense in which he describes it, it is clearly not sufficient for intersubjectivity.

4.1. Footballers and chess players – work, embodiment and shared intentions

Is the physical dimension of the football game crucial to the players' experience of the normative nature of collective action, or could the essential facts of the situation be transferred to a non-physical sphere such as that of a quiz team? In all other respects, the situations are similar – the members of the quiz team and the football team are focused on a common goal, bound by established and accepted rules, and with each individual engaged in more or less the same activity, with some subtle variations in role (the difference between the attacker and the defender, and between the sports specialist and the history specialist). One essential difference between the two scenarios is that Merleau-Ponty's point about the external barriers placed on any physical activity applies only to the football and not the quiz scenario. There remains a physical element to the quiz example that could fall under the heading of Merleau-Ponty's concept of intercorporeity, which is facial expression and gesture. The quiz team whose members are familiar with each other's physicality will come to recognise the subtleties of expression and gesture to signal someone's confidence in a given action or decision, which, depending on the set-up of the particular quiz, could be seen as a skill to gain an advantage which is honed over time until the maximum grip is reached and the symbiosis of their actions reaches perfection.

The situation with the footballers is almost exactly the same. The physicality involved in the quiz players' action is not incidental to the intersubjective action.

At this point, it is worthwhile to ask whether perhaps the terms "intention" and "goal" need to be carefully distinguished, and refer back once more to the notion of basic and higher purposiveness. There is clearly a distinction in everyday speech, as an intention is generally a firm plan to perform a particular action – e.g., I intend to pick up my umbrella before I leave the house in order to serve my *goal* of remaining dry should it rain. Sometimes we use the term "intend" to refer to goals that are very near or achievable. The sentence "I intend to be the President of the United States" sounds somewhat odd unless uttered by someone about to take the oath of office in the next few weeks, or on the brink of being elected. The way Searle uses the term "intention", for example in his discussion of the group of friends rushing to get out of the rain, focuses on simple and straightforward activities that lead to a short-term goal and do not necessarily involve the suppression of one's natural or first-order desires. Indeed, action from this level of basic purposiveness might well be motivated solely by such natural desires, as in Searle's example (Searle 1990). Can we say, then, that sharing collective goals of basic purposiveness does not demonstrate full intersubjectivity and empathy?

Bratman, in Bratman 1993 and elsewhere, provides us with a suggestion using the vocabulary of subplans which have to mesh in particular ways. He uses the vocabulary of "shared" rather than "collective" intentions, which I have regarded as synonyms thus far in this analysis but can clearly be differentiated in an account of meshing subplans. "Shared" seems more appropriate for Bratman's analysis because the intentions or goals produced by meshing subplans results in an analysis which posits common content that directly concerns the social world. According to Bratman, we can share an intention that we wash the dishes if and only if:

1. (a) I intend that we wash the dishes and (b) you intend that we wash the dishes

2. I intend that we wash the dishes in accordance with and because of 1a and 1b, and meshing subplans of 1a and 1b; you intend the same.

3. 1 and 2 are common knowledge.[18]

On my description of basic and higher purposes, washing the dishes would count as a basic purpose subordinate to the higher purpose of, for example, maintaining a pleasant and hygienic living environment. On Bratman's account, subplans in 1a and 1b could be "washing the dishes with Brand A washing-up liquid" and "washing the dishes with Brand B washing-up liquid". Could subplans A and B also be basic purposes serving a higher purpose in the terms described above? What difference would it then make if the higher purpose was not shared? For example if we imagine the context of a newdesk of a newspaper, the higher purposes could be, for one person, to improve sales figures for the week, and, for another, to impress a particular government minister in order to gain an advisory position. The subplans on the level of basic goals could mesh but be serving different higher purposes – the subplans could both mesh in such a way that the basic goal is to write and publish a scandalous story on a political rival to the minister in the second worker's subplan. Depending on all kinds of facts about the particular situation, the shared intention could persist if there were complete common knowledge about everyone's higher purposes. Indeed, it would be perfectly possible for there to be a higher purpose on the part of the newspaper's proprietor that does not overlap, or is even antagonistic towards, the higher purposes of the workers (for example, her intention could be to discredit the government in general for some political purpose). Whilst subplans as they are described by Bratman provide an empirically convincing description on the micro-level, and could also fit in well with a Merleau-Pontian account of skill where close association allows subplans to be carefully balanced in order to achieve mutual satisfaction, they cannot explain why it should be necessary that higher goals and purposes should be shared on the macro-level. In itself this is not an objection to an account that makes use of the concept of subplans, but it directly contradicts what might be seen as Husserl's very promising account of mutual striving which focuses more obviously on complex, higher, long-term shared goals.

[18] See Bratman 1993, p. 106; Bacharach and Tollefsen 2008, p. 32.

Conclusion – The Wider Ontological Framework

In this paper, I have attempted to demonstrate that the way in which joint commitments and goals fit into the general framework of our desires and our Background/idios cosmos is crucial for intersubjective action, and that, for this reason, the phenomenological experience of work is a paradigmatic example of effective intersubjectivity and human interaction. I have suggested, based on Husserl's account of communal striving, that work that is seen by its group workers to have a higher purpose involving the subordination of basic purposes to enable more effective intersubjectivity because of the skills that are developed as a result of such work, and that are used in such work. Perhaps somewhat counter-intuitively, my enquiries suggest that there is no special bonus for intersubjectivity when group workers agree on the higher purpose of their work, that is, what that higher purpose is. In the same vein, it makes no difference to intersubjectivity in working environments whether or not a group of workers is somehow deceived or mistaken about the higher purpose of their work. I have also observed that accounts of intersubjectivity that are relevant to the experience of work strongly tend to support a simulationist account of empathy, where the empathiser experiences the emotions of the Other rather than ascribing them to her on the basis of a theory of mind. I suggest, following Hegel and using an argument from Merleau-Ponty, that individual physical work (or, more specifically, goal-directed activity) can improve one's capacity for intersubjectivity by encouraging the worker to think of themselves as a subject and an object simultaneously, since physical activity brings with it the consciousness of the limits of one's freedom.

As for the comparison between physical and non-physical work, I have examined Merleau-Ponty's account of skill and suggested that it could apply equally to non-physical activities. I suggested also that his account of intercorporeity could function as a paradigm of the intersubjectivity of action, since this phenomenon is most acutely observed when two or more people are engaged in similar physical work together, although not necessarily for the same basic or higher purpose (as in the example of the footballers). I suggest that physical work is therefore more likely than non-physical work to foster solidarity, and that there is also an element in solidarity which concerns hardships experienced by oneself and the Other, making physical work more relevant to the phenomenon than non-physical work. At the same time, since all work by my definition involves the suppression of natural desires, hardship

on some level is always involved in work, despite the "rewards" in terms of higher purpose. In this sense, feelings of solidarity are likely to arise from any type of work.

In general, I found the difference between physical and non-physical group work in terms of fostering intersubjectivity and empathetic understanding to be one of degree rather than form, and maintain that the whole range of intersubjective relationships and empathetic reactions that arise from group work are equally possible in non-physical group work. At the same time, such relationships and reactions are particularly likely to develop in physical group work – indeed, there might also be a biological dimension in terms of mirror neurons, endorphins and lactic acid in the muscles. However, I do not believe that this biological dimension is necessary for the development of intersubjectivity and empathy.

One extremely important question that remains is that of whether these instances of intersubjectivity and empathy must have an ontological basis. I can make only the briefest remarks about what this paper adds to this particular debate here. What I mean by a "monistic ontology" is described in a concise manner by Rolf-Peter Horstmann in a 2006 paper:

> The entirety of actuality must [if we are to accept a monistic ontology] be seen as a single all-comprehending, self-developing rational entity, which achieves knowledge of itself in a spatio-temporal process of realizing its distinctive conceptual determinations. (Horstmann 2006, p. 109)

I can make only the briefest of comments on this topic here, namely that all of the aspects of promising theories of intersubjectivity and empathy as they apply to the world of work have in common a concern with a balance of experience between objectivity and subjectivity. This is certainly not enough in itself for an argument for a monistic ontology, but is perhaps the starting-point of an enquiry into the relationship between the phenomenology of intersubjectivity and the wider ontological framework.

REFERENCES

Bacharach, S., & Tollefsen, D. (2008). Collaborative art and collective intention. In H.-B. Schmid, K. Schulte-Ostermann & N. Psarros (Eds.),

Concepts of Sharedness: Essays on Collective Intentionality, (pp. 21-40). Heusenstamm: Ontos Verlag.

Bratman, M. (1993). Shared intention. *Ethics, 104*(1), 97-113.

Brubaker, D. (2000). Merleau-Ponty's three intertwinings. *The Journal of Value Enquiry, 34*(1), 89-101.

Cornejo, C. (2008). Intersubjectivity as co-phenomenology: From the holism of meaning to the being-in-the-world-with-others. *Integrative Psychological and Behavioral Science, 42*(2), 171-178.

Donohoe, J. (2004). *Husserl on Ethics and Intersubjectivity: From Static to Genetic Phenomenology*. Amherst, NJ: Humanity Books.

Dreyfus, H. (2002). Intelligence without representation – Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences, 1*(1), 367-383.

Drummond, J. (2002). Forms of social unity: Partnership, membership, and citizenship. *Husserl Studies, 18*(2), 141–156.

Flay, J. (1984). *Hegel's Quest for Certainty*. Albany, NY: SUNY Press.

Frankfurt, H. (1988). Freedom of the will and the concept of a person. In H. G. Frankfurt, *The Importance of What We Care About*, (pp. 11-25). Cambridge: Cambridge University Press.

Hadreas, P. (2007). *A Phenomenology of Love and Hate*. Aldershot, UK: Ashgate.

Heal, J. (1995). How to think about thinking. In M. Davies & T. Stone, *Mental Stimulation: Evaluations and Interpretations*, (pp. 33-52). Oxford: Blackwell.

Hegel, G.W.F.(1979). *The Phenomenology of Spirit* (trans. A.V. Miller). Oxford: Oxford University Press

Hegel, G.W.F. (1991). *Elements of the Philosophy of Right*. (tr. by H.B. Nisbet). Cambridge: Cambridge University Press.

Horstmann, R.-P. (2006). Hegel's *Phenomenology of Spirit* as an argument for a monistic ontology. *Inquiry, 49*(1), 103-118.

Husserl, E. (1950). *Cartesianische Meditationen und Pariser Vorträge. (Husserliana I)*. The Hague: Martinus Nijhoff.

Husserl, E. (1973). *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass, Zweiter Teil. (Husserliana XIV)*. The Hague: Martinus Nijhoff.

Husserl, E. (1976). *Die Krisis der Europäischen Wissenschaften und die Transzendentale Phänomenologie: Ein Einleitung in die Phänomenologische Philosophie. (Husserliana VI)*. The Hague: Martinus Nijhoff.

Husserl, E. (1991). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie: Zweites Buch: Phänomenologische Untersuchungen zur Konstitution. (Husserliana IV).* The Hague, Martinus Nijhoff.

Kelly, G. A. (1984). Notes on Hegel's *Lordship and Bondage*. In J. Stewart (Ed.), *The Phenomenology of Spirit Reader*, (pp. 172-191). Albany, NY: SUNY Press.

Kojève, A. (1969). *Introduction to the Reading of Hegel*. New York: Basic Books.

McDowell, J. (2009). *Having the World in View*. Cambridge, MA: Harvard University Press.

Meijers, A. M. (2003). Can collective intentionality be individualized? *American Journal of Economics and Sociology, 62*(1), 167-183.

Merleau-Ponty, M. (1962). *The Phenomenology of Perception*. London: Routledge and Kegan Paul.

Merleau-Ponty, M. (1964). *The Visible and the Invisible*. (tr. by A. Lingis). Evanston, IL: Northwestern University Press.

Reynaert, P. (2001). Intersubjectivity and naturalism – Husserl's Fifth Cartesian Meditation revisited. *Husserl Studies, 17*(3), 207-216.

Scheler, M. (1970). *The Nature of Sympathy*. Hamden, CT: Shoe String Press.

Searle, J. (1990). Collective intentions and actions. In P. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions in Communication*, (pp. 401-415). Cambridge, MA: MIT Press.

Searle, J. (1995). *The Construction of Social Reality*. New York: The Free Press.

Storey, D. (2009). Spirit and/or flesh – Merleau-Ponty's encounter with Hegel. *PhaenEx, 4*(1), 59-83.

Theunissen, M. (1991). The Repressed Intersubjectivity in Hegel's Philosophy of Right. In Cornell, D., Rosenfeld, M. & Carlson, D. G. (Eds.), *Hegel and Legal Theory*, (pp. 3-63). London: Routledge.

Williams, R.(2000). *Hegel's Ethics of Recognition*. Berkeley, CA: University of California Press.

Wood, A. (2006). Fichte's intersubjective I. *Inquiry, 49*(1), 62-79.

Wringe, W. (2003). Simulation, co-cognition, and the attribution of emotional states. *European Journal of Philosophy, 11*(3), 354–374.

Practical Intentionality: A Balance Between Practical and Theoretical Acts[*]

Susi Ferrarello * *
ferrarello.susi@tiscali.it

ABSTRACT

The main purpose of my research is to examine that kind of *intentionality* which underpins every *decision*. By the help of Husserl's phenomenology, I would answer to the following questions: Is every decision determined by a logical reasoning and an axiological evaluation? What do we commonly mean with practical intentionality and how does it affect a *decision-making process*? Usually the idea of intentionality has been linked up to the activity of giving meaning about what one wants to do. I can intend something to acquire information on it and then deed. Although it is simply a matter of opening a window. However, I retain that even in these early stages of the intending, the practical acts convey their own kind of intentionality which could be occasionally covered up by theoretical acts.

1. INTRODUCTION

The main purpose of my research is to examine that kind of *intentionality* which underpins every *decision*. By the help of Husserl's phenomenology, I would answer to the following questions: Is every decision determined by a logical reasoning and an axiological evaluation? Or is it a result of mood and instinct? What exactly flows in a mood? Why are my decisions not always successful?

For instance, when one decides to go out for a walk, he/she could be compelled to make this decision because of a logical inference. I know that outside is a warm and sunny day, so I evaluate the situation in a positive way

* I am grateful to Federico, Valeria and the staff of Humana.Mente for their helpful contribution to this article.

* * University of Rome "La Sapienza"

and I decide to go out and enjoy the pleasant day. Yet the example becomes more complicated when one thinks fit to want something, but his/her body does not seem agree. Namely, I want to get that job, because it is commonly held in great esteem and it is well paid, but once I reached it, I begin to suffer from daily migraines or worse I behave myself so as not to keep it even if I am sure to do my best.

Thence my research aims mostly to a definition as clear as possible of what we commonly mean with practical intentionality and how it affects a *decision-making process*. Usually the idea of intentionality has been linked up to the activity of giving meaning about what one wants to do. I can intend something to acquire information on it and then deed. Although it is simply a matter of opening a window. However, I retain that even in these early stages of the intending, the practical acts[1] convey their own kind of intentionality which could be occasionally covered up by theoretical acts.[2]

To carry out this analysis I will rely on Husserl's *Analyses Concerning Passive Synthesis* (Husserl 1966[3]) and *Active Synthesis* (Husserl 2000[4]). Following Husserl's path I should be able to draw an analysis of will conducive to put in evidence both subordination and superiority of willing respect to knowing. In fact, by Husserl's former works (Husserl 1975[5], Husserl 1984[6] and Husserl 1988[7]), I will emphasize why he conceives will subordinated to the logical reason. On the contrary, quoting some passages of his latter studies (particularly Husserl 2004), I will be able to display how Husserl combines his former with his latter idea of will and accordingly his theory of knowledge with his idea of decision-making process.

2. Affective and Signitive Acts

In this paragraph I will try to provide with a tentative definition of what Husserl

[1] Although the rank of practical acts is quite wide, here I will refer to Husserlian definition of practical acts as it is given in his lectures on ethics (Husserl, 1914). The practical acts encompass the acts of will, evaluations, emotions, instincts, feelings, sentiments.

[2] In the class of theoretical acts Husserl usually puts the acts of judgment and knowledge.

[3] From now on: Hua XI.

[4] From now on: Hua XXXI.

[5] From now on: Hua XVIII.

[6] From now on: Hua XIX.

[7] From now on: Hua XXVIII.

means by *intentionality* and which kind of rationality lays at the basis of this idea.

Intentionality is mainly the term by which phenomenologists describe the movement of consciousness to mean something. In *Logical Investigation*, Husserl mainly uses the term of intentionality in the meaning of *Bedeutungsintention* (Benoist 2004 and Souche-Dauges 1998). As Føllesdal remarks, the Husserlian perspective consists in the description of *Richtung* not of the object toward which the act is aimed, but of a certain structure of consciousness (Føllesdal 1990).

> Das determinierende Beiwort *intentional* nennt den gemeinsamen Wesenscharakter der abzugrenzenden Erlebnisklasse, die Eigenheit der Intention, das sich in der Weise der Vorstellung oder in einer irgend analogen Weise auf Gegenständliches Beziehen. Als kürzeren Ausdruck werden wir [...], das Wort Akt gebrauchen. (Hua XIX, p. 406)

The *Erlebnisse* of pure consciousness are intentional acts that for economy's sake can be called 'acts'. It is relevant to note here that we are going to handle this term in the sense of *mental act*, that is the act as we perceive it (Hua XIX, p. 406).[8] In fact in this quest, we have no interest in describing the action as a bodily movement or an external intended action, rather we want to figure out how the mental act of a practical decision comes to realization. In the fifth *Logical Investigation*, Husserl employs two definitions to classify all the acts of consciousness; he writes that all the psychic phenomena are characterized by an intentional reference and «sie entweder Vorstellungen sind oder auf Vorstellungen als ihrer Grundlage beruhenit» (Hua XIX, p. 406). Therefore the *Erlebnisse* of consciousness can be either acts *(reel)* or non acts *(real)*. All the acts are intentional; on the other hand, non acts are not intentional because they do not relate to any represented object. They are tied to the psychological sphere of consciousness. «Dass nicht alle Erlebnisse intentionale sind, zeigen die Empfindungen und Empfindungskomplexionen» (Hua XIX, p. 382). Sensations arise without an object being represented or sensed. Accordingly these are not properly intentional. Nevertheless acts of feeling can also be taken up into the range of intentional acts: «Sie alle 'verdanken' ihre intentionale Beziehung gewissen ihnen unterliegenden Vorstellung» (Hua XIX, p. 404).

[8] Husserl borrows the distinction between mental and physical acts on the basis of an inner and external perception from Brentano's *Psychologie* (1874).

> Ein Kentaurenkampf, den ich mir in einem Bilde oder in der Phantasie vorstelle, 'erregt' ebenso mein Wohlgefallen wie eine schöne Landschaft der Wirklichkeit, und wenn ich die letztere auch psychophysisch als reale Ursache für den in mir seelisch erwirkten Zustand des Wohlgefallens auffasse [...]. Das Wohlgefälligsein, bzw. das Wohlgefallenempfinden 'gehört' zu dieses Landschaft nicht als physikalischer Realität [...], sondern in dem hier fraglichen Aktbewusstsein gehört es zu ihr als so und so erscheinender evtl. auch so und so beurteilter, an dies oder jenes erinnernder usw. als solche ‚fordert', ‚weckt ' sie dergleichen Gefühle. (Hua XIX, p. 405)

Here it can be pointed up the influence of Brentanian thought. In fact a feeling of pleasure may be intentional whenever it is provided with the representation of the object.

> Findet man eine Schwierigkeit darin, dass nicht jedes Begehren eine bewusste Beziehung auf ein Begehrtes zu fordern scheine, da wir doch oft von einem dunkle Lagen und Drängen bewegt und einem unvorgestellten Endziel zugetrieben werden; und weist man zumal auf die weite Sphäre der natürlichen Instinkte hin, denen mindestens ursprünglich die bewusste Zielvorstellung mangle, so würden wir antworten: Entweder es liegen hierbei bloße Empfindungen vor [...], also Erlebnisse, die wirklich der intentionalen Beziehung ermangeln und daher auch dem wesentlichen Charakter des intentionalen Begehrens gattungsfremd sind. Oder wir sagen: es handle sich zwar um intentionale Erlebnisse, jedoch um solche, die als unbestimmt gerichtete Intentionen charakterisiert sind.
> [...]
> Die Freude ist nicht ein konkreter Akt für sich und das Urteil ein daneben liegender Akt, sondern *das Urteil ist der fundierende Akt für die Freude*, *es bestimmt ihren Inhalt* [...], denn ohne solche Fundierung kann Freude überhaupt nicht sein. (Hua XIX, p. 405, *my emphasis*)

Thus, judgment is always an ultimate act with respect to an act of feeling, because it gives a meaning about which we can feel the sentiment. Joy could be an intentional act only when it relies on the epistemological contents given by the logical reason. We can feel joy just after we know at what we are rejoicing. The content *(Inhalt)* is determined by the judgment.

Nevertheless, also theoretical acts, like those of judgment, can determine their contents by the tools of intuition and perception. They are able to collect all the data which are going to be represented. Every intuitive act is an objectifying act and it encompasses the act of perception. In the sixth *Logical Investigation* Husserl seems to construe intuition as a distinct and particular

property of perception and vice-versa. Intuition is a sort of perception of the universal and then a way of perceiving, which is exploited in order to account for the fullness of meaning, the truthfulness of our perception and the possibility of its representation (Hua XIX, pp. 64-84 *passim*). At large, it is posed on the same stage of perception, even if it seems to work just from the inside of consciousness.

> Die Anschauung als Perzeption [...] – gleichgültig ob sie kategorial oder sensual, ob sie adäquat oder inadäquat ist – wird in Gegensatz gebracht zum bloßen Denken als dem bloßen signifikativen Meinen. (Hua XIX/2, p. 731)

Both intuition and perception can be addressed to an ideal or empirical object which could even be not respondent to reality. Yet both contribute to the effectiveness of any intentional or objectifying act by the meaningfulness fulfilment of signitive acts (which could be considered empty boxes until then).

Therefore intentionality could be defined as the skill of mental acts to be directed to an object. Objectifying acts are for Husserl «vorstellig machende Akte»: the acts which make present the intentional object for the consciousness, the acts which institute the intentional relation between consciousness and the object. They do this job also for non-objectifying acts (hence also for affective acts). Objectifying acts are both signitive acts (judgment and representation) and intuitive acts (external and internal perception, eidetic intuition, imagination, remembering, empathy, etc.). Accordingly objectifying acts are both acts of meaning and thinking (signitive acts) and acts of intuition. Yet, affective acts are not fully acts as they require the empty boxes of signitive acts to express at all their intention or even to exist. As Husserl wrote about the joy, it calls for the judgment in order to be an effective act. Without the judgment, it could not have its object on which it operates.

> Das determinierende Beiwort *intentional* nennt den gemeinsamen Wesenscharakter der abzugrenzenden Erlebnisklasse, die Eigenheit der Intention, das sich in der Weise der Vorstellung oder in einer irgend analogen Weise auf Gegenständliches beziehen. (Hua XIX, p. 46)

In this sense the act of feeling has its essence *(Wesenscharakter)*, but it is founded on a judgment because it needs the predicative voice of logical acts (or the boxes of signitive acts) to determine its object. This is still more evident in Husserl's ethical lectures of 1914, when he talks about a *Verflechtung* between practical and logical acts to explain completely how an affective

intention works.

> In ihr <Parallelismus> drückt sich eine gewisse Wesensverflechtung des doxischen Bewusstseins mit dem Gemütbewusstsein und so jedem Bewusstsein überhaupt aus, dergemäss jedes Stellungnehmen, jedes Schön- oder Gut-Werten *apriori* in ein urteilendes Stellungnehmen umgewandelt werden kann. (Hua XXVIII, p. 63)

The will could not know what to want if it has not a box where to put its feeling. Therefore it is necessary to display the process of practical intentionality by the device of interlacing and parallelism between affective and signitive acts.

3. Practical Intentionality

Now if the signitive acts are fundamental to express an affective act and if they represent or judge just what I can already represent or know, how do I make a decision on what I just feel but not understand? How could I make a decision if I do not know all that I am feeling? Still, if intentionality is the skill of mental acts to be directed to an object and if this skill is balanced on the complicated relationship between objectifying (signitive and intuitive acts) and not objectifying acts (affective and conative acts, i.e., practical acts), how could I justify these mental acts or their objects? How could I fill the lack of the object of practical acts and their inability to create 'new boxes'? To answer these questions, I should deepen the understanding of how Husserl define *will* in his former and latter studies.[9]

In Husserl's former work, namely in the lectures on ethics (Husserl 1988)[10] will is defined as one of the several regions of consciousness and it holds a prominent function to connect consciousness with the outside world (Hua XXVIII, p. 59). In his *Husserls Phänomenologie des Willens*, Melle (1992) helps us to sketch out the main influences on Husserl's idea of will. Namely, he refers to the work of James (1950) and Ehrenfels (1887), since the former retains that the main characteristic of will is the attention and its *fiat*, the 'act of mental consent'. The latter construes the act of will not as a founding act, because it is just a pretension *(Forderung)* of something and thus it needs the representative and theoretical acts. As it concerns the present research, it

[9] I will refer to Hua XIX, XXVIII and Hua XI, XXI, XXXVII.

[10] These lectures are the result of the ethical researches carried out by Husserl since 1902.

could be fruitful to pay attention to the influence exerted by both philosophers. In fact, the will could be depicted not only as a region of consciousness but also as a kind of intentionality which underpins every action even if it is always interlaced with its founding signitive acts.

In 1914 Husserl conceived will as a way of consciousness' being which needed the represented objects to exist. In other words, its object is a represented content which is already explained, at least formally, by signitive acts. The distinctive characteristic of will is not the intended object, but its "*fiat!*" (Hua XXVIII, p. 107), i.e., that kind of power addressed toward the object. Will adds something to the structure of intentionality, since it is the motor of any act and it yields new reality. As a matter of fact, every act of will modifies, in a certain way, reality or leads it toward new directions. As Husserl remarks, the thesis of will *(Willensthesis)* is mainly a position of realization and creation which is interwoven with the position of theoretical acts (in virtue of their 'Allwirksamkeit' or predicative voice, see Hua XXVIII, p. 58). When one wants to go for a walk, one should know before what a walk means or rather what he needs, then one decides to go. During the walk new situations could be generated by his/her decisions.

Another element, which should be emphasized in this analysis, is the axiological component. According to Husserl of 1914 every decision arises from an epistemological intention which is evaluated by the axiological reason (Hua XXVIII, pp. 70-71). The inference of any decision should be drawn by the evaluation of what it is given. One should decide what to do, after having understood what to do and evaluated what is the best for him/her. Consequently, citing Husserl's words

> Das Alles ist Sache der vernuenftigen Konsequenz. Aber solche Konsequenz verbindet auch das intellective Gebiet mit den Gemütsgebieten; theoretische und wertende Vernunft sind miteinander überall verflochten. (Hua XXVIII, p. 72)

Thus, will makes its decisions on the predicative voice of theoretical acts (i.e., its meanings) and on the evaluation of the axiological acts. Consequently, will is a rational region of consciousness which depends strictly on theoretical acts to interact with reality.

3.1 Second version of practical intentionality

In Husserl's *Analyses concerning Passive and Active Synthesis* (Hua XI, XXXI)

the definition of will is not exactly the same, as though, like Peucker (2008) claimed, Husserl will keep in a certain way the former view. In this latter version of practical intentionality he will also seek to give an answer to the following questions: what happens when one begins to pay his/her attention to an object, such as a dawn or a laughing child? Why is one attracted to an object more than others? There could be a kind of objects which pertains specifically to the sphere of practical reason? Then does practical intentionality exist?

At first glance, I believe that the beginning of knowing is practical. Effectively it is quite easy to show that the intention to know is driven by the choice to know and intend the object. I want to focus on a dialogue between two friends of mine instead of the noise of a barking dog (or vice-versa), because I decided so. Husserl explains this attitude by the key concepts of *affection (Affektion)* and *attention (Aufmerksamkeit)*. In the *Analyses Concerning Passive Synthesis*, Husserl explains that the affection is the first striving which exerts its influence on attention and then on intention (Hua XI, p. 152). Affection arises in contrast with what is used to perceive – «Affektion ist also Kontrast» (Hua XI, p. 149). We can take the example of the walk given in the first paragraph: one is walking in a sunny day and a sudden outburst could interrupt all that he/she was perceiving. This event will change the focus of his/her perception. Affections are located exactly in this feeling of contrast which arises from what one is used to perceive. Affections are driven also by the mixture of attention and interest. The act-motivating passive sphere does not only consist of merely neutral presentations, but it is rather penetrated by elementary strivings and feelings which carry evaluative features. Affection is a sort of 'emotional' emergence which comes out so strongly that one is compelled to move his attention to another field of perception (Hua XI, p. 149). It drives the direction of evaluation, representation and attention. Another example could be the one given by Husserl himself, I am listening to a music and everything around me is stimulating my senses, when a sudden loud noise attracts my attention. I am affected by that noise and my previous affections has been interrupted by this new event. In this case, I have been stricken by an event which is not consistent with all that I used to perceive. Consequently, I can claim that «Affektives Relief» (Hua XI, p. 168) characterizes the passive foundation of what we can call a *practical*

intentionality.[11] These acts are always combined with certain evaluative features that are given in feelings, and only these qualitative differences inside the passive sphere of affection can explain why the ego turns toward an affection but not toward another. Therefore, attention (and then perception) is a tension got in motion by affection. «Das Affektion zur Aufmerksamkeit, zur Erfassung, Kenntnisnahme, Explikation sich auswirkt» (Hua XI, p. 151). The attention is that form of tension which allows the practical and passive intentionality to become active and operating. If these blind drives work themselves out, they neither involve the activity of a genuine act of will nor the ego. Husserl describes these subjective occurrences as the intentionality of drives *(Triebintentionalität)* and he even calls them a very "low form of the will" or "passivity of will". As he wrote:

> jedes [...] *ego cogito* ist an die Voraussetzung gebunden, dass vorher das Ich affiziert wurde, das sagt, das vorher eine passive Intentionalität, in der das Ich noch nicht waltet, einen Gegenstand konstituiert hat, von dem aus der Ichpol affiziert und zum actus bestimmt worden ist. (Hua XI, p. 209)

The antichamber (*«Vorzimmer»*, Hua XI p. 166) of every decision is a combination of passive affections which strike the attention and its activity.[12] Thus the intention to know or to give a meaning about what I am living is always preceded by a practical intention. The interlacing between these two positions has not necessarily a theoretical prevailing thesis.

Moreover, in *Analyses concerning Active Synthesis*, Husserl emphasizes the role played by the *Willensintention* and complains about its misinterpretation. The consistency between attention and affection is the first step of any practical or theoretical decision. «Der Wille ist kein bloßes Begehren; er gehört in die allgemeiner Sphäre der reinen Aktivität» (Hua XXXI, p. 10). Will represents a very activity which involves aware and discretionary acts.

> Es will mir immermehr scheinen, dass Wille nicht eine eigene Weise des Bewusstseins ist, sondern eine besondere und höhere Form der Aktivität, die unter gewissen Wesensbedingungen, die in vorausgesetzten Objektivierungen

[11] See Hua XXXVII, pp. 339-340: «Allem Triebmäßigen, mich affektiv Motivierenden oder zu motivieren Tendierenden schleudere ich mein ewiges Nein entgegen. Die willensbestimmende Kraft aller passiven Motive durchstreiche ich. [...] Triebe dürfen mich nur motivieren, wenn ich sie an der Leine habe, wenn ich ihnen ihre Funktion und den Rahmen ihrer Funktion vorzeichne».

[12] See expressions as «eine niedere Form des Willens», «Willenspassivität» (Ms. M III 3 102f).

und Fühlungen liegen, überall auftreten kann. (Hua XXXI, p. 10)

The will represents a particular and superior aspect of the rational activity of consciousness and it can 'come into play' under certain conditions. These conditions coincide with those of 1914, namely with those objects of meaning that the acts have to assume before making a decision. The activity of will is considered as a particular and superior kind of rational action, in virtue of the key role played by affection and attention with driving the interest to perceive something. Attention is a *positive* feeling, or better it is an act that *makes* the interest real. This feeling makes a simple act of perception, an act of interest. Attention is in fact the tension 'to be' in the things that we perceive. As it was in the *Psychology* of James, taken by Husserl as a model to his investigation[13], the tension of attention is the main instrument to fix the direction of will. It can change perception in interest and interest in will. As a matter of fact, it adds to the interest 'the tension' which unifies the *ego* to the object of perception and will to its productive characteristic (namely, the skill to modify reality and to yield new reality by its *fiat*). For instance when one is listening to music, his/her perception is focused at all on that. A sudden outburst moves the focus of attention on itself. The affection is driving the attention to change the aim of his/her interest. Then, the simple act of perception becomes an act of interest and active will.

Differently from the lectures on ethics (1988), now the affective acts, particularly the volitional ones, are not totally subordinated to the representations of signitive acts to be effective, because the very first beginning of their intention is a passive and instinctive strive. In fact, Husserl wrote in his manuscript that the reason is always a practical reason and it is servant of will (Ms, E III, 7, 85). There is an intertwined coexistence between practical and logical position within practical intentionality. Will is a primitive form of action; it is *'superior* and *particular*' because it is at the basis of all kinds of acts, also of the logical acts. Indeed, the first step of any acquaintance is not a real form of knowledge but a 'will to know'. The true knowledge consists in the productive action of the ego. The logical reason can be really

[13] In 1891/92 Husserl took a class on psychology and on that occasion he read for the first time the *Principles*. In may 1894 he came back on *Principles*, while he was working on his logic and its elementary concepts and he praised Jamesian effort of "depsychologizing psychology". At that time he had planned to publish a series of articles in the *Philosophishe Monatshefte*, but he published only the first and decided to wait to see what James had done, before publishing the others. The next article is probably his *Psychological Studies for Elementary Logic*.

directed to the knowledge, only if the will realizes itself in the will of doing. As Nam-In Lee wrote, every kind of intentionality should be regarded as a practical intentionality (Lee 2000), because every act of consciousness is always a practical act. Also Hart remarked that only by volitional acts it is possible to put in light all the contents of the other kinds of reason (Hart 1992). In fact, the theoretical reason does not understand its representations until they do not reflect on them. And even the acts of reflection are the result of the *fiat* of will. As Husserl wrote, every act is an act of will (Ms A V, 22, 5). The predicative activity of logical reason is still necessary to give voice to the reality we know, but differently from the lectures of 1914, now (lectures of 1920) the process of communication between consciousness and the world is not due to theoretical but to volitional acts. Even if the former are still essential to make possible the expression of what we know, the latter are the starting point of every logical act.

4. How Do We Make a Decision?

To elucidate all that we acquired until now, we can claim that the practical intentionality is explained on the complicated balance with signitive and affective acts. Both are equally relevant in the making-decision process *(even a decision to know!)*. I make a decision because: 1) an affection moves my attention toward a certain object, 2) I can represent what the object is and fulfill my representation with meaning thanks to my intuition and perception, 3) I can evaluate how much this object is important to me. Yet, if my intention remains just a passive level, the second point is not always needed. In fact, I can decide instinctively what to do just on a dim sensation without knowing exactly why I decided so.

Generally, I perceive a dawn and not a barking dog because at that time, my attention is attracted to the dawn and I want to admire it. If I want to understand why one has a particular affection rather than another, I have to keep in mind that affections are not just neutral theoretical data. Conversely they work on the emergency of contrast in our habits to perceive. The emotional objects toward which the acts are directed, are epistemological and represented data lived in a practical way (a passive, affective or instinctive way). In 1920 Husserl is not so far from what he stated in 1914. He simply added new elements to explain the practical components of intentionality. In his

lectures course on ethics from the 1920s he explicitly says that

> Wertende Akte und Willensakte sind in Erkenntnisakten, eventuell schon in bestimmenden Urteilen, fundiert [...]. Was ist nicht mindest vorstelle, kann ich nie werten.

Consequently the founding of the volitional acts has to be understood against the background of this passive sphere in which theoretical, emotional, instinctive and drive-related tendencies are already mutually intertwined. The acts of willing are not based on mere presentations and some higher ordered feelings, but rather on the dynamic processes of the passive subjective life in which a separation of independent spheres of acts would make no sense. As a matter of fact, the feeling consciousness presupposes a cognitive act, while the volitional consciousness in turn presupposes the feeling (Hua XXXVII, p. 274).

Therefore, when one seeks to decide what to know, one is equally led by theoretical and practical intentions which are intertwined in the same act of intending the outside world. In the making-decision process, sometimes it happens that what I perceived from my practical instincts (i.e., from all the practical components of my will) is put aside by my theoretical acts of knowledge. Consequently, it would take place a processes of detachment. Therefore, a successful decision should be based on a perfect consistency between theoretical and volitional acts. The reasons provided by knowledge become usually stronger than those given originally by practical acts. It could even happen that I forget all that I originally felt about a certain object, a job for instance, because I assign higher values to what theoretical acts say. Then I decide to follow what is *logical* without reflecting on all the components of my will. Nevertheless, in this complicated balance between logical and practical acts it remains an open and unsettled issue how my feelings could be represented if I do not know well all that I am feeling. I can make my decision just on what I already know about my practical acts (feelings, instincts, sensations) because of the cognitive limits of signitive acts. This kind of communication could be a term of a new philosophical research.

REFERENCES

Benoist, J. (2004). La fenomenologia e i limiti dell'oggettivazione: il problema degli atti non obiettivanti. In B. Centi & G. Gigliotti (Eds.), F*enomenologia della Ragione Pratica*, (pp. 153-174). Napoli: Bibliopolis.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Leipzig: Duncker & Humblot.

Von Ehrenfels, C. (1887). *Über fühlen und Wollen*. Wien: Carl Gerold & Sohn.

Føllesdal, D. (1990). Noema and meaning in Husserl. *Philosophy and Phenomenological Research*, L, 263-271.

Hart, J. (1992). *The Person and the Common Life*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Husserl, E. (1966). *Analysen zur passiven Synthesis. Aus Vorlesungs- und Forschungsmanuskripten, 1918-1926* .M. Fleischer (Ed.). The Hague, Netherlands: Martinus Nijhoff.

Husserl, E. (1975). *Logische Untersuchungen. Erster Teil. Prolegomena zur reinen Logik*. Text der und der 2. Auflage, Halle: 1900; rev. ed. 1913, E. Holstein (Ed.). The Hague, Netherlands: Martinus Nijhoff.

Husserl, E. (1984). *Logische Untersuchungen. Zweiter Teil. Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*, Halle: 1901; rev. ed. 1922, U. Panzer (Ed.). The Hague, Netherlands: Martinus Nijhoff.

Husserl, E. (1988). *Vorlesungen über Ethik und Wertlehre, 1908-1914*. U. Melle (Ed.). The Hague, Netherlands: Kluwer Academic Publishers.

Husserl, E. (2000). *Aktive Synthesen: Aus der Vorlesung Transzendentale Logik 1920/21 Ergänzungsband zu Analysen zur passiven Synthesis*. R. Breeur (Ed.). The Hague, Netherlands: Kluwer Academic Publishers.

Husserl, E. (2004). *Einleitung in die Ethik, 1920/1924*. H. Peucker (Ed.). Dordrecht/Boston/London: Kluwer Academic Publishers.

James, J. (1950). *Principles of Mental Psychology*, vol. 1. New York: Dover Publications.

Lee, N.-I. (2000). Practical Intentionality and Transcendental Phenomenology as a Practical Philosophy. *Husserl Studies, 17*(1), 49-63.

Melle, U. (1992). Phänomenologie des Willens. *Tijdschrift voor Filosofie, 54*, 280-304.

Peucker, H. (2008). From logic to person. *Review of Metaphysics, 62*(2), 307-319.

Souche-Dauges, D. (1998). *Le développement de l'intentionnalité chez Husserl*. Paris: Vrin.

Is Balancing Emblematic of Action? Two or Three Pointers from Reid and Peirce

David Vender *
dvender@utas.edu.au

ABSTRACT

Defining actions in contradistinction to mere happenings runs into the problem of specifying the role of the agent and separating what the agent does from what they exploit or suffer. Traditionally these problems have been approached by starting with a simple act, such as an incidental movement, and considering causality, or by seeking to elucidate the connection between the act and the agent's intentions or reasons. It is suggested here that a promising approach is to shift attention from 'simple' movements and start instead by exploring the general character of acquired skills. Balancing the body is one such skill and serves here as an exemplar. Some remarks made by Reid on balance are used in a Peircean framework for perception to suggest that, at least for humans, an action is always the performance of an acquired skill. Also, while action is constitutive of perception, bodily perception is the basis of action, providing in a feeling of ownership direct knowledge of an asymmetric opposition between the agent and the world.

ACTION, INTENTION, FREEDOM

Attempts to understand human action have often been framed in the context of the problem of free will. This relies on an analysis of some behaviors of agents, particularly those thought to require the kinds of motivation informed by reason, custom and moral purpose. Distinguishing which aspects of these behaviors qualify them as acts – in contrast to those behaviors which lie beyond the agent's control or influence – runs into diverse difficulties, particularly if the explanatory efforts also take up the task of naturalizing agency and the agent's intentions and purposes.

* School of Philosophy, University of Tasmania

Many problems may be thought of as arising from difficulties in separating the agent, with their incidental motivation and *arbitrating and arbitrary* judgment, from the presumably orderly processes which underlie the expression of the agent's intentions.

The object of an act may well be to cause some change in the external world, something like the shifting of a stone, which is well described by elementary physics. But it may also be to move a part of the agent's body, and this is only partially described in physiological models of moving organisms.[1] The object may even be to suppress an unwelcome memory, and this type of act is even less well understood. In each case the difficulties turn on the role of the individual agent, and drawing a boundary line between the agent and the realm of effects at the skin, or perhaps the periphery of the central nervous system, is rarely satisfactory.

In preference to exploring how this difficulty in defining agency and action operates in previous analyses of action, such as those of Davidson or Frankfurt, the present paper seeks to identify a kind of behavior which might be taken as emblematic of action. Examining this behavior may not escape all the traditional difficulties of analyzing action, but it may lead to some insight into how these difficulties arise and which directions of inquiry seem promising for their resolution.

The approach taken is to consider balance, which underlies human perceptual and bodily orientation as well as the active maintenance of posture, and through that the performance of every movement, including locomotion. I draw on some ideas from Thomas Reid[2] and Charles Peirce in order to do justice to balancing and its characteristic phenomenology.[3]

[1] The notion that muscular motion is now well understood by analogy with cybernetically sophisticated robots is quickly dispelled if one considers historically the issue of the operation of the heart, as done by Thomas Fuchs (2001).

[2] Particularly the remarks in the 1795 essay called *Of Muscular Motion in the Human Body*, published in Wood 1995.

[3] The sense in which 'phenomenology' is intended here is not the phenomenology of modern philosophical schools, but the kind of Baconian bringing together of sundry but relevant facts and characteristics which enables systematic inquiry by making the first steps towards a general description possible. Cf. Peirce 5.37; in common with the secondary literature, the Collected Papers of Peirce are cited by volume and paragraph number.

Skilled Action

Recent work on the explanation of action is largely confined to the tradition of treating actions as events and exploring how causes and reasons might play into these events as well as the complications introduced into the conception of agency by morality, determinism and the specification of intention (Sandis 2009). Human intentions might be described in terms of goals and aims, or purposes, and evaluated in connection with reasons and expectations. Perceptual knowledge, direct and mediated, plays a central part in all such descriptions, but the kind of knowledge amenable to declarative expression as objective fact is insufficient. The self-reflective awareness of some striving, the apprehension that it is I who is acting in at least a subjectively voluntary and deliberate way also appears to be necessary. We do not have to insist that this awareness actually dominates the agent's concerns, only that if the relevant action is suitably attended to, these aspects of its character are evident to the agent. In this sense, the signature of action can be said to be the agent's ownership of it.

Focusing on the awareness of what the agent feels to be doing in acting shifts the inquiry away from an examination of the reasons which the agent may cite in justifying or explaining their actions. While this may seem to make an analysis of the agent's reasons more, not less, difficult – and it may be protested that intention without reasons is no intention – it may actually help to clarify important matters. It helps by bringing into sharper relief the factors more immediate to the agent's act. These are best considered first, leaving conceptual re-descriptions of the act waiting until the primary characteristics of an act are agreed upon.

The contrast here is not a difference between the how and why of an action. It is more a matter of neglecting, at least initially, elaborated narratives justifying what happened or even constructing *post hoc* explanations of it. These narratives and explanations have a more theoretical character, and their development must logically follow from an appraisal of those data which ultimately validate these theories. These data are initially given in how the agent perceives their own doing.

It is a truism that an individual agent rarely understands much of what constitutes their act. Not only do chains of Why? questions retreat unstoppably in manifold directions, the consequences of even trivial acts can hardly be followed out fully, so in large part they remain obscure and unheeded.

Intention and result also usually match in only a rough-and-ready manner. Furthermore, in performing even the most routine movements (or, if we insist on separating them out, mental acts) the agent's body and brain are implicated in exquisitely complex processes, most of which the agent has no inkling of. Even now our most advanced scientific inferences regarding these processes are sketchy at best, and demonstrably incomplete. But such knowledge is not needed to make the agent feel empowered. Finally, even the stories that agents tell themselves and others about the reasons for their own behavior are subject to the limitations of self-knowledge and the knowledge of the context in which an act arises.

Given all this, it would seem best to begin with the direct perception of what we are doing when we act, rather than trying to impose a conceptual system on inherently complex acts, whose available justification is schematic.

Doubts concerning a correct identification of the intentions informing an act, and an appreciation of its full complexity, point quite precisely in the direction where we should seek the kind of actions most suitable for an initial attack on the general problem. Confidence that we are acting successfully in accord with our intention and purpose is characteristic of skills which we learn and finesse through patient repetition and practice.[4]

Rather than considering isolated acts such as an incidental movement of an arm (which can be justified by an endless list of unrelated intentions) it is better to start with routines which have a constant character. The movements used for walking and running are not always or even often precisely the same, but walking is easily distinguished from running or standing still, and whatever additional reasons may apply in specific cases, the immediate intention informing those movements is to walk or run, and success or failure are relatively easy to judge. It is of no concern if the penultimate reason for our skilful movement just now is rarely (if ever) successful performance of the movement itself. What matters is that our ownership of the movements as agents or actors is uncontroversial, the immediate reasons for them are well defined, and criteria for success are clear.

[4] Skills have recently been considered in the context of causal theories of action by Clarke 2010, who seeks to quell worries that skills undermine causal accounts by trying to extend these to cover skills such as shaving and dancing. The present work differs not only in leaving causality aside until some of the basic phenomenology of skilled behavior is clear, but also in attempting to dispel the notion that «skilled activity differs in important ways from many of the stock examples that are employed in action theory, such as raising one's arm» (Clarke 2010, p. 523).

Balance in the Human Animal

What has just been said about walking and running points to the fundamental importance of balance and orientation in routine human actions. Humans are unusual animals in that it takes them an inordinately long time to acquire the habits needed for even minimal locomotion. The high degree of plasticity and incomplete development of the human brain at birth are important factors in explaining why humans are so slow at first. The usual posture they finally adopt is also precarious, requiring constant monitoring and feedback for its maintenance. The bipedal stance may be useful in minimizing the moment of inertia around the vertical axis – facilitating a quick turnaround – but it comes at a price.

An understanding of orientation and balance and how these are exploited in holding posture and getting about is quite recent. We might even speculate that before it could be developed, Newtonian physics had to displace Aristotelian paradigms. Its development was also hampered by the fact that the sensory part of the story is complex and well hidden.

Information on the dynamical variables needed to maintain balance is obtained through the use of several groups of organs, among which the most important is the vestibular apparatus of the inner ear. The semicircular canals and utricles which are parts of these organs contain mechanoreceptors which are used to detect rotational and linear accelerations, and these are instrumental in orienting and stabilizing the head in relation to the reference frame of the earth. It is easy to see how important this information is for the perception of the location and motions of physical objects.[5]

The significance and functioning of the vestibular apparatus has been clarified only recently.[6] However, even before the role of vestibular functioning in providing the basic dynamical information needed for orientation and movement became known, the Scottish philosopher Thomas Reid said some remarkable things about balance.

[5] These comments should not be misconstrued as a suggestion that fully functioning vestibular organs are necessary for balance. These organs are grossly impaired in some deaf individuals who can nevertheless attain balance by means of other organs, using various receptors in the muscles and joints, particularly in the neck.

[6] See Howard and Templeton 1966. Early research focused on vertigo, motion sickness and nystagmus. Wade 2000 presents the early history. Recent developments have been reviewed by Angelaki and Cullen 2008.

Reid was led to consider bodily sensations by his epistemological scheme, in which subjective sensations function as signs for the real qualities of bodies.[7] Since we are manifestly able to move our bodies in space, we must be able to *feel* bodily motions and exertions in order to control our limbs effectively, and we perceive the direction of the gravitational force, or whichever resultant force acts when we are accelerated bodily through space. Reid spoke about our balancing not only in a way which appeals to common sense, but noticed some characteristics which we should never lose sight of.

The first of these is that Reid prioritises perception in action, noting that:

> There are however many voluntary Motions in which some previous Perception of the Understanding is necessary to direct us to the Motion which the occasion requires. (Wood 1995, p. 110)

Reid is primarily concerned with how active agents use the muscles, but he does not make the problematic move in insisting that we must at each moment be conscious of the muscular movement, strain, position and whatever else is required to specify the initial conditions for and the performance of a particular act. He recognizes that much of this may be subliminal or unattended, and by his epistemological scheme is led to search for sensations which in the normal course of action are «absolutely unheeded», as he puts it in a related context (Reid 2000, p. 82).

The second important characteristic of maintaining our posture is that it requires unceasing effort:

> Although all voluntary Motion is performed by the Contraction of Muscles, we must not from that conclude that when no Motion is willed, the Muscles are inactive. *The Exertion of Muscles is no less necessary to rest than to Motion. In every position of the Body excepting perhaps that of lying prone.* (Wood 1995, p. 112, *emphasis in original*)

The third important characteristic is that balance is not something that we learn once and for all. It must be continuously cultivated and can even be improved:

> When we observe with what ease, and Grace those Motions are performed by those who are expert, and compare them with the Laws of Motion, we must be convinced that this Sense by which we perceive the least deviation of the Body

[7] A valuable introductory account is provided by Wolterstorff 2001.

> from its Ballance, may by Use be brought to a degree of Accuracy which is hardly to be observed in any of our other Senses. (Wood 1995, p. 110)

Finally, the fourth important point is that the actions underlying balance do not require explicitly formulated purposes to be meaningful – we might say that verbally or conceptually elaborated explanations and justifications can in some sense remain only implicit in actions. Reid does not say this explicitly, but he focuses on sensations and feelings in perception, and notes that balancing is of immediate concern to the pre-verbal infant:

> This sense of Ballance may be seen in a Child of two or three Months old. If sitting upon ones knee he begins to tumble, he immediately starts & endeavours to recover himself; But it is greatly improved by Use, in every Employment that requires its exercise; [...] This sense of our Ballance is produced not onely by the impression made by the power of gravity but by any other Force which endangers the Ballance. (Wood 1995, p. 111, spelling original)

As already stated, Reid thought about balancing well before the functioning of the vestibular apparatus was clarified. Modern research has revealed that this set of organs does have the most significant position among the organs we use to perceive the downward direction and rotational motions of the head. Not only are these dynamic data crucial for orderly movement, they play a fundamental role in perceptual development, and it is not too much to say that our ability to see objects located in and moving through space is founded on the integration of information on dynamical variables mainly from vestibular receptors with light signals detected by the retina. The vestibular organs mature early – even before brain structure develops fully – and the chief perceptual learning tasks for the infant appear to be to integrate visual and vestibular signals so that they can see like an adult, while separating their sensations into visual, auditory, olfactory and other streams.[8]

It is remarkable that even here Reid, who was keenly interested in medicine and surgery and a careful observer of children, has something interesting to say. Although he felt obliged to maintain that our perceptions of primary qualities such as extension and hardness were original and unlearned, he left room in his epistemology for acquired perceptions. He does insist, against Berkeley, that we see depth immediately, yet he notes that:

[8] Empirical work on infant development supporting these assertions is presented by Maurer and Maurer 1988.

> From the time that children begin to use their hands, nature directs them to handle every thing over and over, to look at it while they handle it, and to put it in various positions, and at various distances from the eye. [...] It is this childish employment that enables them to make the proper use of their eyes. They are thereby every day acquiring habits of perception, which are of greater importance than any thing we can teach them. (Reid 2000, p. 201)

Balancing bodily members is the first step in the control of movement, developing even before the upright stance is achieved. Lifting and turning the head are important in the infant's first efforts. Once control of movement is adequate, control not only determines the character of all our movements, it is also fundamental for not moving. Keeping still and maintaining a particular orientation or attitude is the basic requirement for seeing remote objects, indeed for all visual perceptions, which we control instrumentally by turning the head, directing the eyes and then *keeping the gaze directed*. The link between the eyes and the vestibular apparatus is so strong that compensatory eye movements which preserve clear vision while the head is moving exhibit the character of reflexes. In humans, however, this vestibulo-ocular 'reflex' is learned, plastic, and adaptive when the apparent motion of visible objects is artificially manipulated (Benson 1982).

It would seem that a problem of the genesis of agency arises here. There are two reasons why we should not get distracted by it in considering action. The first is that understanding agency and understanding the genesis of agency can, at least to some extent, be separated. An analogous situation exists in the domain of language. This too is a problem of agency since the question being asked is when we first decided to associate arbitrary signifiers with reasonably constant meanings. The origin of language is a formidable puzzle, but the structure and continuing development of languages can be studied profitably without solving it. It is just so with action.

The second reason why the question of origins is not as acute as it may appear is that habits do not get started from scratch. The awakening infant is not faced with a perceptual nothingness, a kind of blank screen in a stationary void. Their body is already structured and their field of experience is pregnant with possibilities of action. The development of agency is not the initiation of movements from a dead stillness. It is the gradual bringing of order and expectation into the operations of an animated body, and taking control of pre-existing motions and adapting them creatively for invented purposes. How

adults do this can be considered without fully understanding how infants get started, although mimesis is evidently a key ingredient for both.

To sum up, the actions we perform depend on balancing the body and making efforts and as such are combinations of learned skilled acts. Action has a recursive structure. We do not assemble any movement 'from scratch', but try to adapt previously performed actions to the problem at hand, and develop these adaptations by comparing our intentions and expectations to the effects of the action. Perception, memory and imagination are the three cognitive pillars of this process, and balancing is the central activity which allows the agent to pursue their particular goals – both perceptual and operational – as a physically effective participant in the real world. Instead of now leaving this central activity aside in favor of considering abstract notions of causality or the conceptual structure of how specific acts are justified, it is better to remain with balancing in order to explore how we perceive our own effectiveness in acting. Peirce is a valuable guide in these matters.

Perception in Doubt, Effort, Habit and Skill

Reid's epistemology was based on a dualism of mind and body, and while he was an enthusiastic proponent of science and of efforts to naturalise the mind, he resolved the problem of relating subjective experience to objective reality by an appeal to an order preordained by God. This explanation carries little weight now, and dualism is seen to underlie some difficult problems in naturalizing subjective states.

A fresh approach to these problems can be found in the ideas of Charles Sanders Peirce. Not only is Peirce one of the foremost authorities on the methodology of modern science, he was also thoroughly anti-Cartesian in his epistemology and in his metaphysical speculations. However, his opposition to dualism did not turn him towards materialism. On the contrary, he felt it necessary to formulate new categories which could support a unified theoretical framework not just for psychology, but also for language and logic.

It is not necessary to enter into the technical details of Peirce's theory of signs in order to describe action from his perspective. It will be sufficient to consider his categories of firstness, secondness and thirdness – which he never

tired of describing and explaining – and how they relate to subjective experience.[9]

Action is at the heart of Peirce's version of pragmatism and only a sketch of how he explained its characteristics is attempted here.[10] For Peirce, pervasive doubt in the style of Descartes is a methodological hoax, a pretense at best. Actual doubting is a felt irritation at the failure of expectation, present mainly when our habitual actions do not adequately meet their imagined ends. Actions are informed by beliefs, and the «essence of belief is the establishment of a habit; and different beliefs are distinguished by the different modes of action to which they give rise» (Peirce 5.398).

To dispel any impression that this may be related to behaviorism, we only need to turn to the primacy of thinking in what Peirce calls belief and action. «The soul and meaning of thought [...] can never be made to direct itself toward anything but the production of belief» (Peirce 5.396).

> As it appeases the irritation of doubt, which is the motive for thinking, thought relaxes, and comes to rest for a moment when belief is reached. But, since belief is a rule for action, the application of which involves further doubt and further thought, at the same time that it is a stopping-place, it is also a new starting-place for thought. (Peirce 5.397)

Thinking (and in general all inference and cognition) is a process which takes time. If we wish to comprehend what it is, we must examine what we can become aware of when we are actually thinking.

Peirce analyzes this self-reflective awareness into three subjectively distinguishable categories of conscious experience which, while they are always all present when suitably attended to, modify the character of our awareness as one or another predominates. These categories can most briefly be characterized as a pure quality (e.g., redness) for firstness, a dual opposition or relation for secondness, and a threefold relation for thirdness. The last has the general nature of the sign and it informs our awareness when we find some symbol or experience meaningful. For Peirce these categories are not invented descriptions of subjective episodes but «modes of being» which he sought to

[9] Peirce brings particular expertise to this topic too, since he made a seminal contribution to the development of psychophysics by developing measurement techniques and introducing statistical methods.

[10] A fuller treatment can be found in Potter 1997, where what I wish to call action is more often called habit, and the role of the classical normative sciences of esthetics, ethics and logic (as Peirce understood these) is explained.

validate and apply through scientific, logical and philosophical explorations (Peirce 8.328-332). The most important category to consider first in connection with balance is secondness.

Among varied illustrations of secondness, the one relevant for us is physical effort:

> Standing on the outside of a door that is slightly ajar, you put your hand upon the knob to open and enter it. You experience an unseen, silent resistance. You put your shoulder against the door and, gathering your forces, put forth a tremendous effort. Effort supposes resistance. Where there is no effort there is no resistance, where there is no resistance there is no effort either in this world or any of the worlds of possibility. (Peirce 1.320)

What is explained here applies precisely to balancing. The sensory and motor aspects are inseparable. This does not mean that we immediately lose orientation and perspective if we lie down and relax, since perceptual and cognitive habits can persist against neglect for some time, but it does mean that prolonged isolation from opportunities to refresh dynamical perceptions through active efforts must be expected to lead to such loss. In balancing we are participants in a supra-individual order, but this order has to be actively – i.e., voluntarily – explored by the participant. As embodied knowers we are not spectators, but actors. Now Peirce insists that secondness is irreducible:

> You have a sense of resistance and at the same time a sense of effort. There can be no resistance without effort; there can be no effort without resistance. They are only two ways of describing the same experience. It is a double consciousness. We become aware of ourself in becoming aware of the not-self. The waking state is a consciousness of reaction; and as the consciousness *itself* is two-sided, so it has also two varieties; namely, action, where our modification of other things is more prominent than their reaction on us, and perception, where their effect on us is overwhelmingly greater than our effect on them. (Peirce 1.324)

The notion of cause expresses secondness, as does any constraint. The flow of time, in how the past is expressed in the present, does also. The contrast between sensing (feeling) and will is in how we trace the antecedents. If these are internal we are agents, while:

> In sense, the antecedent events are not within us; and besides, the object of which we form a perception [...] remains unaffected. Consequently, we say that we are patients, not agents. In the idea of reality, Secondness is predominant;

> for the real is that which insists upon forcing its way to recognition as something *other* than the mind's creation. (Peirce 1.325)

What we normally call sensing is thus for Peirce secondness as much as doing is. Even in the simplest perceptions, such as the awareness of a color, secondness intrudes. Not necessarily, to be sure, through the awareness of any effort, but through the externality of the quality itself. This is sometimes expressed by calling the color 'given', but Peirce also emphasizes the fact that color is not perceived as color *simpliciter*, in a kind of anoesis, but as located and spread out (Peirce 1.313n1).

We have noted that balancing is the foundation of perspective and orientation. It is also, through the vestibular and other organs, the basis of the *directed spatiality* which we call spatial awareness. Objects are not merely in space, they lie in a particular direction and occupy a definite location. Sense impressions are not simply extended, or distant, they arise *from* a specific somewhere relative to the perceiver's viewpoint.

The complexity of our direct experience in the course of the development of skills, indeed in any *doing*, has been noticed, and much can be gained in realizing that our awareness is mischaracterized if it is thought to consist simply of attention directed sequentially to this or that thing or feeling.[11] But it is not enough to admit that awareness is rarely if ever unitary, and to convert the passive perceiver into an actor by making it dual. What is still missing is thirdness, which expresses the fact that the objects of our consciousness are *all*, at least to some extent, meaningful. This is to say that in recognizing something, we comprehend at least minimally *what kind* it is or, equivalently, what might or might not be done about it.

Thinking in Action

In common with other philosophers, Thomas Reid's theory of perception was a sign theory (Clark 2007, ch. 10). Simple unitary experiences, such as the impression of a vivid color or the sound of a bell, act as signs. These signs coupled with certain judgments inform us about objects and events in a way analogous to how we grasp the meaning of words. The knowledge acquired this way is superior to the mere enjoyment of sensations, and Reid distinguished

[11] See Polanyi 1969 and Sennett 2008.

sensation from perception, claiming that when we perceive we not only understand the significance of particular sensations, but we are assured of the relevant object's independent existence.

Reid did not go into much detail on how a sign acquires meaning and how it is understood. The use of the analogy between perception and comprehension takes for granted our familiarity with language in order to illuminate perception. If one wishes to go further than naïve views on language, what is needed is a theory of signs.

In his attempts to formulate a general theory and classification of signs, Peirce came to believe that for something to be a sign three elements had to come into relation. This threefold unity could not be reduced to a set of dual relations and still keep its functionality. The simplest illustration of this interdependence may be gathered by considering that a symbol cannot have a meaning until it is properly embedded in a system: a group of letters cannot be a word until it has a place in a language. A dual association, such as between a written symbol and a sound, is only a code, not a symbolism.

Peirce presents a barrage of explanations and arguments to make himself understood, but rehearsing any of these would divert us too far from action. Suffice to say that the development of the idea of thirdness may come directly from logical considerations, from an examination of inference, and anyone wishing to argue that thirdness is reducible needs to do so by (irreducibly) bringing three terms together – hence the would-be reductionist cannot practise what they preach.[12]

While the theory of signs developed by Peirce is complex and the terminology he used to classify signs mind-boggling in its unfamiliarity, the motivation for developing it can readily be understood when we consider routine actions such as balancing and keeping still. Just as the human awareness rarely if ever rests in firstness, so the experience of secondness is not a simple feeling of dual consciousness in which efforts strive blindly against opposition. Our efforts are directed and we attach at least a minimal significance to them (Peirce 1.532). Without this significance or meaning we may fail to identify the feelings and sensations experienced, and tend not to even perceive them.

[12] Cf., «When people ask me to prove a proposition in philosophy I am often obliged to reply that it is a corollary from the logic of relatives» (Peirce 1.629).

Our intelligence is an intelligence that deals with signs. In striving to do anything, what is present to our reflection is not a bare feeling, but an effort which has this (rather than some other, or no) direction, as well as some significance and expectation indissolubly bound to it.

In characterizing an intelligence which deals with signs, it is important not to restrict the meaning of 'sign' to lexical constructs. Signs are available to the human intelligence even before the mastery of language, and anything at all can serve as a sign to this intelligence. We are primarily not language users, but thinkers, and while using language is perhaps the most efficient form of thinking for some purposes, it is not exclusive.

Peirce described our intelligence in a telling manner as «a "scientific" intelligence, that is to say, [...] an intelligence capable of learning by experience» (Peirce 2.227). Not only is this directly relevant to the exercise of skills and to experiencing «genuine doubt» (Peirce 5.443) – which to Peirce is a truly affective state – it also allows a 'scientific' intelligence to be pre-verbal.[13] The only prerequisite is that this intelligence is an active, thinking one, i.e., one judging expectation against result and modifying its future actions and expectations in the process. As is evident from the quotation on page 260 above, for Peirce this process is the essence of thinking.

Understanding thinking in this way advises the adoption of a very inclusive conception of inference and indeed:

> When Peirce speaks of an "inference," he means *any* cognitive activity whatever, not merely conscious abstract thought. Specifically, he includes perceptual knowledge and even subconscious mental activity. (Davis 1972, p. 9)

There is on this account no fundamental difference between a syllogism expressing clear conceptual relations and worked through explicitly from premises to conclusion, and the routines implicit in perceptual habits or in acting generally. Perceiving and acting are subsumed into forms of inference, and a categorical difference between knowing how and knowing that becomes untenable. As Peirce puts it: «To act intelligently and to see intelligently become at bottom one» (Peirce 7.562).

This all inclusive nature of what are taken to be thinking and inference might provoke the worry that the generality of this theory makes an account of action unusably vague. If we cannot even keep practical skills separate from book knowledge – two accomplishments which are clearly not interchangeable

[13] Cf. Peirce 5.227-235.

– how can we hope to formulate a clear difference between action and mere behavior? Some brief comments can be offered to suggest that this worry is unfounded.

If the difference between action and behavior is sought in the explicability of actions in terms of the agent's reasons, the recursive complexity of reasons advises that we are not in any position to simply match actions (classified perhaps as various movements) with reasons. It must be sufficient that suitable reasons *can* be given and that we are convinced, on investigating the concrete case, that the agent performed the act. The role of the agent presupposes effort, but for the agent to be appropriately involved in the act any effort must be directed and as such grounded in those skills which underlie orientation and balance. This is why balancing can serve as exemplary of the 'simplest' kind of act.

It is not essential for the agent to attend to any particular aspect of their performance – they are usually captivated by the goal anyway. However, it is important that the skills relevant to the act have been acquired by the individual in question in the inferential cycle starting from expectation and going through doubting, thinking, and settling on belief, as Peirce explained it. It is only the adequate repetition of this cycle which can furnish a movement with a felt significance, and it should come as no surprise that two of the earliest verbal expressions of infants are those of satisfaction with something well done and disapproval at some action whose result did not meet expectation (Gopnik *et al.* 1999).

The only kind of action we can perform is one which is constituted from a combination of learned skilled acts. While the underlying skill is the signature of action, it is still quite possible that there is no rule which can be formally applied to differentiate between action and behavior in any specific instance. It is even likely that the distinction may need to be drawn differently for various acts or for different agents. Much depends on what the individual agent has acquired some measure of control over. This imperative to remain in the concrete might be an impediment to formal theory, but it is not an impasse in practical life where common sense counsels that the most effective agents learn by doing.

The distinction between behavior and action need not be abandoned, but a spectrum is revealed ranging from bodily processes which we have never thought to influence or master, all the way to what Peirce called conduct, which is «action that is self-controlled, i.e., controlled by adequate deliberation»

(Peirce 8.322). Although we judge children differently from adults, and also subject the unintended consequences of adult actions to the arbitration of judgment, wherever we recognize that some skill or mastery has been acquired by the individual, there we accept that the individual is acting.

Broken Symmetries

The perspective arising from the participation in a dynamical order by balancing makes all our actions necessarily directed. This directedness is part of the meaning of all our movements, and even those acts which are normally spoken of as if no movements were involved – mental acts such as imagining and thinking – turn out on close inspection to be closely related to physical movement.[14]

The directionality of our movements as well as the spatial content of our perceptual states presupposes an asymmetry between a here and a there. It makes all the difference in the world if something moves from here to there or vice versa. There are also such differences between what it takes to move upwards voluntarily – as in standing up, jumping or climbing – and downwards – as in falling or crouching – that it would take very peculiar circumstances for us to confuse them.[15]

The particular perspective of our experience is evidently consistent with the spatial order in which our physical body exists, and we cannot literally move in a direction orthogonal to the three axes defining up-down, left-right, and forward-back. However, there would seem to be no logical necessity in a universe to have a certain spatial or temporal order, so the embodiment we enjoy as biological organisms on earth can be at least speculatively taken as contingent. This raises the question of what may be the minimal requirements for an intelligence to be active.

The idea that perceiving is possible without embodiment in three dimensions was already considered by Reid (2000, pp. 108-112). Elaborating on a hint from Berkeley, Reid imagined a race of spirits who see but cannot

[14] For the intimate relation of thinking to what may seem trivial or superfluous movements see McNeill 2005, who explores the deep connections between gestures and verbal expression. Reid believed that we share the language of gesture with the animals.

[15] This is in contrast with the perfect symmetry of the action and reaction pair in Newtonian physics.

touch. These Idomenians lack the notion of a third dimension and to them objects occluded by nearer bodies are theorized to be 'overcome', but both objects must occupy the same space since those occluded have nowhere to hide.[16] Reid used this fable in developing a non-Euclidean spherical geometry for visible (depthless) objects (Grandi 2005). However, even for these hypothetical beings a perspective enabling rotations is necessary so the symmetry between here and there is (dynamically) broken.

The dynamical asymmetries just mentioned arise from embodiment, which allows us to participate in the physical world. Still, the asymmetry inherent in the directional perspectivity of this participation is neither the same nor likely to be sufficient for us to feel that it is we who are acting. There would seem to be an experiential difference between perceiving that our body is moving in a particular direction and the knowledge that we are striving in that direction. It is this asymmetry between effort and resistance that Peirce pointed to in describing the dual consciousness, and the duality comes not from a simple opposition, or even from the opposition of two directions, but from the fact that we feel ourselves to be the owners of one side of the opposition of forces, of the balance.

There are thus at least two asymmetries operating in physical action, and if one asks about the necessity of embodiment for action, what is being asked includes asking how dynamical participation relates to the ownership felt when we act. It may be true that, as a matter of fact, these asymmetries are inseparable in our course of life. But it is difficult to decide on this basis alone whether they must be inseparable. If they can be separated then it would seem that it is the apprehension of ownership that is necessary, while how this ownership is exercised, be it through directed movement or through some other perhaps difficult to imagine process, is unessential.

Saying that a *feeling* of ownership is essential in acting is not the same as claiming that we must be aware of our body or in any particular affective state while performing an act. It is often said that in acting it is precisely these bodily feelings and states which we neglect, and when we balance we generally do so unthinkingly.[17] When action is considered as a skilled performance, however,

[16] It is interesting that the ontological persistence of occluded or hidden objects is a kind of discovery for infants, and this relates to the popularity of 'peek-a-boo' games. See Gopnik *et al.* 1999.

[17] An argument against the necessity of explicit bodily knowing (performative or affective) in some specific acts has been given by Young 2004. It is based on pathological cases, so its impact on a description of action in general is limited.

what we are momentarily aware of in acting proves to be consistent with the phenomenology of craftsmanship.[18] In exercising a skill we are intent on the end result and, having mastered the skill, can afford to neglect attending to what the performance requires of us. Yet, just as reasons can be supplied after the fact, we can rehearse our movements and choose to pay closer attention to them and our ownership of them whenever the need arises.

Conclusion

In seeking to understand action, the first task is to identify those actions which are typical and can serve as exemplars of human agency. The next important step is to trace how these are developed and cultured, since human actions are best characterized as performances of acquired skills. In examining how we *perceive ourselves to be acting*, the asymmetry which Peirce defined as dual consciousness would seem to be a fundamental requirement, but moving becomes acting only when an intelligence which deals with signs thinks through its actions and modifies them to meet expectations.

Following Reid's indications, I have suggested that balancing is emblematic of action. Not only is it a cultured skill, it serves as the basis for the whole variety of human actions, including those highly cultivated acts which follow from deliberation and are explicitly justified by causal explanations and reasons. We do not have to be fully aware of our contribution for something to count as an act, or be able to justify it rationally, but we must be able to adapt our efforts to the momentary situation which we perceive ourselves to be in, so that our expectations have some hope of being met.

References

Angelaki, D. E., & Cullen, K. E. (2008). Vestibular system: The many facets of a multimodal sense. *Annual Review of Neuroscience, 31*, 125-150.

[18] Craftsmanship is considered in detail by Richard Sennett (2008). Michael Polanyi has also made valuable remarks on these topics, even explaining how active doing informs book knowledge in the acquisition of expertise, and formulating the idea of 'tacit knowing' to characterize how bodily knowledge underlies meaningful activity in Polanyi 1969, part 3.

Benson, A. J. (1982). The vestibular sensory system. In H. B. Barlow & J. D. Mollon (Eds.), *The Senses*, (pp. 333-368). Cambridge: Cambridge University Press.

Clark, S. (2007). *Vanities of the Eye: Vision in Early Modern European Culture*. Oxford: Oxford University Press.

Clarke, R. (2010). Skilled activity and the causal theory of action. *Philosophy and Phenomenological Research, 80*(3), 523-550.

Davis, W. H. (1972). *Peirce's Epistemology*. The Hague: Martinus Nijhoff.

Fuchs, T. (2001). *Mechanisation of the Heart: Harvey and Descartes*. Rochester, NY: University of Rochester Press.

Gopnik, A., Meltzoff, A. & Kuhl, P. (1999). *How Babies Think: The Science of Childhood*. London: Phoenix.

Grandi, G. B. (2005). Thomas Reid's geometry of visibles and the parallel postulate. *Studies in History and Philosophy of Science, 36*(1), 79-103.

Howard, I. P., & Templeton, W. B. (1966). *Human Spatial Orientation*. London: John Wiley & Sons.

Maurer, D. & Maurer, C. (1988). *The World of the Newborn*. New York: Basic Books.

McNeill, D. (2005). *Gesture & Thought*. Chicago: University of Chicago Press.

Peirce, C. S. (1931-1958). *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

Polanyi, M. (1969). *Knowing and Being*. Chicago: University of Chicago Press.

Potter, V. G. (1997). *Charles S. Peirce on Norms & Ideals*. New York: Fordham University Press.

Reid, T. (2000). *An Inquiry into the Human Mind on the Principles of Common Sense*. University Park, PA: Pennsylvania State University Press.

Sandis, C. (Ed.) (2009). *New Essays on the Explanation of Action*. Hartfordshire: Palgrave Macmillan.

Sennett, R. (2008). *The Craftsman*. London: Penguin Books.

Wade, N. J. (2000). William Charles Wells (1757–1817) and vestibular research before Purkinje and Flourens. *Journal of Vestibular Research, 10*(3), 127-137.

Wolterstorff, N. (2001). *Thomas Reid and the Story of Epistemology*. Cambridge: Cambridge University Press.

Wood, P. (Ed.) (1995). *Thomas Reid on the Animate Creation: Papers Relating to the Life Sciences*. Edinburgh: Edinburgh University Press.

Young, G. (2004). Bodily knowing: Re-thinking our understanding of procedural knowledge. *Philosophical Explorations, 7*(1), 37-54.

Book Review

Effective Intentions: The Power of Conscious Will

Alfred R. Mele
Oxford University Press, Oxford, 2009

Marco Fenici *
fenici@unisi.it

Mele's book is a concise analysis of much research in neurophysiology and neuroscience – starting from the pioneering works of Libet (Libet 1985, Libet *et al.* 1983) – claiming that, although we conceive ourselves as free agents with the power to influence our behaviour by our volitions, free will as well as the causal power of conscious intentions are illusions. Against this claim, Mele argues that it depends on a naïve picture of human agency, thus, it disappears if we develop a sophisticated framework about the explanation of action. When understood according to this framework, empirical data is open to alternative interpretations, and it does not warrant the illusion thesis anymore.

Here is a summary of the book. In Chapter 1, Mele introduces the basic psychological notions involved in ordinary explanations of action. The largest part of the book is then devoted to an extended analysis of many empirical results. In Chapters from 2 to 4, Mele challenges Libet's (1985, 2004) claim that the brain "decides" to initiate actions prior to subjective awareness of the decision. In Chapter 5, Mele argues that the phenomena grounding Wegner's (2002) thesis about the illusion of free will are instead consistent with the causal relevance of intentions in the production of action. In Chapter 6, Mele argues that Lau, Rogers and Passingham (Lau *et al.* 2007) have not shown that conscious proximal intentions emerge too late to be among the causes of corresponding intentional actions. The last two Chapters focus positive evidence to Mele's thesis that there are effective intentions, that is, «intentions that issue in corresponding actions» (p. vii). Hence, Chapter 7 reports

* University of Siena

empirical data supporting the causal role of conscious decision to the production of action. Finally, in Chapter 8, Mele discusses which empirical discovery would persuade him of the truth of the thesis about the illusion of free will and of the causal power of conscious intentions.

Mele sometimes discusses much specific technicalities, an analysis of which is beyond the scope of this review. Herein, I will just attempt to re-compose the book's general strategy by putting together Mele's several–and sometimes fragmented–discussion about contemporary research. Mele aims to depict a mature framework where the concept of intention may be defined in accordance to the thesis that intentions play a causal role in the production of intentional action. He sketches this framework mostly in Chapter 1 by largely referring to his previous works (Mele 1992, 2003, 2007).

Mele attempts to precisely identify the concept of intention as it appears in the discussion about the illusion of free will. According to him, intentions are «*executive* attitudes toward plans» (p. 6). He distinguishes *occurrent* from *standing* intentions – which are dispositions to have corresponding occurrent intentions. Furthermore, he also distinguishes *distal* intentions – that is, intentions which are for the non-immediate future – from *proximal* intentions – that is, intentions to do something in the very moment. He thus explains he will limit his analysis to occurrent proximal intentions – from here on, just "intentions" – because empirical investigation almost exclusively focused their causal role with respect to intentional behavior.

According to Mele, there are two ways – not mutually exclusive – for an intention to *A* to be an occurrent intention at that time:

> One way is for it to be suitably at work at that time in producing relevant intentional actions or in producing items appropriate for the production of relevant intentional actions; the other is, roughly, for it to be a conscious intention at that time, provided that the intention is not wholly constituted by a disposition to have occurrent intentions to *A*. (p. 4)

Thus, Mele rejects the idea that all intentions must be conscious. In order for an intention to be an occurrent intention, it is sufficient "for it to be suitably at work at that time in producing relevant intentional actions or in producing items appropriate for the production of relevant intentional actions". That is, although Mele concedes that some intentions are conscious, awareness is not a necessary characteristic of all of them. Instead, we may identify intentions by their effect – i.e., intentional action. This is a fundamental point to Mele's

general analysis, as it is the key concept to understand how he will later reject the thesis about the illusion of free will and of the causal power of intentions.

Finally, in Chapter 1, Mele also distinguishes intentions from desires, the function of which is to help to produce occurrent intentions. Someone who has a desire may still be deliberating about whether to follow it or not for action. Instead, intentions are more connected to intentional action than corresponding desires. Still, they are also different from practical decisions to do something, in that they may come to be without being formed in acts of deciding.

As I have already remarked, a discussion of the many technical points of contention in the book is beyond the scope of this review. Let me just show how Mele applies his analysis of the notion of intention to one of the most popular experiments leading to claims about the illusion of free will. Libet (1985) wired experimental subjects' with electroencephalogram (EEG) and asked them to flex their wrist at an arbitrary moment. He measured both the shift in the readiness potentials (RPs) in the EEG tracing anticipating the muscle contraction and the time at which the subjects first became aware of their decision to flex. He found that RPs manifested a reliable change 550 ms before subjects begun to flex their wrist, while subjects declared on average to have made the decision to flex only 350 ms before they started flexing. Therefore, Libet claimed, the flexing was triggered by the RP-shift before subjects became aware of their intention to flex. He concluded that intentions are an echo of the brain activity, but that they do not have the power to influence people's decisions.

In discussing Libet's experiment, Mele shows that the experimental data does not warrant that the measured RP-shift stands for subjects' intention to flex their wrist. On the one hand, the experiment does not demonstrate that the RP-shift necessarily triggers the flexing reaction: «'whenever you wiggle your finger, signal *S* appears a second before you wiggle it' does not entail 'whenever signal *S* appears, you wiggle your finger a second later'» (p. 81). On the other, Mele reports much empirical evidence showing that «it is much more likely that what emerges around – 550 ms is a potential cause of a proximal intention or decision than a proximal intention or decision itself» (p. 51), and that it may more accurately characterised in the terms of «urges to (prepare to) flex soon, brain events suitable for being proximal causal contributors to such urges, motor preparation, and motor imagery» (p. 56).

If Libet was wrong in interpreting his experimental data, Mele argues, this is because he relied on what it is «a popular folk theory about intentions or the folk concept of intention, not empirical considerations» (p. 37). Such a folk theory mistakes intentions for conscious intentions. However, as I have noted above, Mele have argued that intentions are not necessarily conscious. Given that, it is still possible for the subjects' conscious experience of their intention to flex to appear later than the RP-shift without we are forced to claim that intentions play no causal role with respect to action. In fact,

> A subject's wanting to flex soon and his experience of wanting to flex soon are not the same thing. So to grant that a subject's flex soon experience of wanting to flex soon is not a cause of his flexing is not to grant that his wanting to flex soon also is not a cause of his flexing. My flipping a light switch—not my *experience* of flipping it—is a cause of the light going on. Analogously, a subject's wanting to flex soon may be a cause of his flexing even if his experience of wanting to flex soon is not. (pp. 32-33)

Therefore, even if Libet were correct about the average time of initial awareness, existing evidence does not warrant his conclusion.

Instead, Libet's experimental data is compatible with the claim that intentions have a causal power in determining intentional behavior. In the light of both the positive evidence attesting the causal role of intentions reported in Chapter 7, and his detailed analysis of the variety of experimental settings leading neuroscientists to claim that free will is an illusion, Mele concludes that a sophisticated analysis of the concept of intention allows one not to exclude that conscious intentions do play a causal role in the production of action:

> Conceived of as essentially supernatural, effective intentions and decisions and the power of conscious will have a ghost of a chance—or, more aptly, a ghost's chance—of existing. Conceived of more naturally, their being every bit as real as you and I are is consistent with the scientific findings examined in this book. (p. 160)

In conclusion, *Effective Intentions* is a nice example of philosophical sensibility applied to scientific research, and it is recommended to both neuroscientists and philosophers of the cognitive sciences. Mele's careful examination of the current debate in psychology and neuroscience about the illusion of free will and the causal efficacy of intentions makes the book a fundamental reading for anyone interested to the topic. However, it should be noted that *Effective Intentions* is all but introductory. Despite the broad scope

and interest of the issues discussed, and despite the remarkable stylistic concision, both the technicality of the analysis of contemporary experiments and the detail of the theses discussed all concur to make the book hard to non-specialists.

Furthermore, concision sometimes is not a merit. Mele's argument that intentions may effectively issue in intentional action strongly depends on their broad interpretation as executive attitudes toward plans. However, this interpretation is as important to the general economy of the argument as much as it is not almost theoretically supported. It may be possible that, having addressed the issue in many of his previous works, Mele did not felt the urge to provide his reader with more details. But it is equally undeniable that the reader would have been more convinced by the whole discussion if such an important piece of the puzzle had been more carefully considered.

REFERENCES

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience*, *19*(1), 81-90.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*(04), 529-539.

Libet, B. (2004). *Mind Time: The Temporal Factor in Consciousness*. Cambridge, MA: Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain: A Journal of Neurology*, *106* (3), 623-642.

Mele, A. R. (1992). *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.

Mele, A. R. (2003). *Motivation and Agency*. New York: Oxford University.

Mele, A. R. (2007). Persisting Intentions. *Nous*, *41*(4), 735-757.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Book Review

The Actor's Brain: Exploring the Cognitive Neuroscience of Free Will

Sean A. Spence
Oxford University Press, Oxford, 2009

Roberto Di Letizia *
r.diletizia@tin.it

Sean Spence's *The Actor's Brain* sees free will through the wide-angle lens of cognitive neuroscience by furnishing the readers with a terrific amount of evidences from neuroscience. Spence's work does not offer only a close examination of the recent studies on volition. In chapter 10, he proposes a new overview on the volitional control as a result of the empirical studies quoted in the previous chapters. The human capacity for volition is presented as a multidimensional space subject to multiple constraints. The volitional control is represented by the 'human response space', the range of behavioral responses that the agent is enabled to perform. 'Human response space' is set by multiple constrains (i.e., factors which determine the boundaries of the human response space), as result human freedom is not a *binary* property, something that humans have or do not have, but a *scalable* property, something that humans have *more* or *less* depending on these constrains. Accordingly, the human response space can be expanded or contracted by changing these constrains. Both internal and external to the subject, these constrains are (i) *anatomical* (chs. 2, 4, 6); (ii) *physiological* (ch. 4); (iii) *neurochemical* (ch. 4); (iv) *psychological* (chs. 2, 8, 7); (v) *emotional* (ch. 9); (vi) *social* (ch. 9); (vii) *genetic* (ch. 9). These constrains are not static as they may be altered in different manners in order to "sculpt" the response space (e.g., drug therapies may potentially restore the response space).Throughout the whole book, Spence explores and examines these constrains.

* University of Salento

In chapter 1, he focuses on the journey of the motor signal, which allows the subject to move the right index, through the *central nervous system* and the *peripheral nervous system*.

In chapter 2, Spence considers the 'anterior' frontal lobes in order to walk backwards towards the initiation of the action. Voluntary behavior is the result of the integrated work of (1) dorsolateral prefrontal cortex (DLPFC), involved in the 'self-generation' of the action and the planning of action of a response, (2) orbitofrontal prefrontal cortex (OFC), implicated in relating the relative 'reward value' to objects or targets towards which the action is directed, and (3) frontopolar cortex (BA 10), involved in planning an alternative response to that programmed by DLPFC. DLPFC and BA 10 plan two different alternative responses, whereas the preference is determined by OFC which attributes 'value' to these perceived behavioral alternatives. Finally, premotor cortex (PMC) has the role in determining the 'script', the 'pattern' of motor events', that the motor cortices may be subsequently called upon to execute.

In chapter 3, Spence faces the timing of volition. Indeed, conscious awareness of acting seems to arise later than the onset of the motor programming and the content of motor programming. These findings suggest that the intention of acting is subject to a double 'delay'. First, the intention of acting precedes our awareness of movement onset. However, the onset of motor programming precedes the finalization of the content of such motor programming. Second, the intentional act is temporally related to the late RP (namely *readiness potential*, the brain's electric activity related to voluntary action), whereas the onset of motor programming is temporally linked to the early RP.

In chapter 4, the main issue pertains to how the brain initiates, modulates, and terminates action, in the absence of a central controller. At a neuroanatomical level, several brain regions are involved in volitional behavior, especially five basal ganglia-thalamo-cortical 'circuits' – re-entrant loops in which information is recurrently re-cycled, in trajectories that are circular – that contribute to volitional control in several manners (e.g., suppression or execution of finely tuned and overly learned motor routines, motor skill acquisition, emotional behavior). At the neurochemical level, volitional behavior is analyzed in terms of neurotransmitters (i.e., dopamine, serotonin, noradrenaline, acetylcholine) whose different levels of distribution may affect both higher and lower aspects of volition. At a cognitive level, Spence follows Tim Shillice's model of volition's cognitive architectures. According to this

model, the human executive system is composed by a lower and a higher system. The lower system performs the routine, automated, and stereotypical behaviors by means of *schemata* – overly learned and simple motor routines which are automated and triggered by cues in our external environment. The higher system (or, executive system) performs consciously planned and spontaneous novel behavior by means of a *'Supervisory Attentional System'* (SAS).

In chapter 5, Spence focuses upon abnormalities of volitional experience, with particular regard to those instances when human agents may be deprived of both their motor control and their sense of agency. Relatively complex behaviors may arise unbidden (e.g., anarchic hands, namely limbs that 'will not do' what their owners 'wish them to do') or under the 'influence' of 'external forces' (e.g., a patient with schizophrenia moves her hand but *feels* as if she is subject to the play of 'cosmic strings'). According to Spence, the organic causes of these diseases are structural and functional abnormalities located in several distributed brain regions, and seem to impair agency via two mechanisms: (i) a disinhibition of 'lower' motor centres giving rise to relatively stereotypic and contextually inappropriate motor routines (e.g., anarchic hands); (ii) a disturbance in the perception of voluntary movement (e.g., alien agency).

Chapter 6 is dedicated to avolition, the apparent *absence* of voluntary behaviors. Avolition is present in schizophrenic patients as they exhibit limited behavioral repertoire and a poor responsiveness to their environment. The poverty of the behavioral repertoire indicates that in avolition the prefrontal and anterior cingulate should be implicated in some way. Indeed, avolitional patients exhibit greater prefrontal lobes deficit, whether in terms of 'function' or 'structure'. According to Spence, avolitional syndromes may emerge when the executive system is impaired (e.g., by genetic factors impacting the dopaminergic and glutamatergic systems), so that the agent's behavioral repertoire is limited to the performances of the subordinate slave system.

Chapter 7 faces a volitional disorder: hysteria. Hysteric patients exhibit unusual, but purposeful, behaviors ('motor hysteria'), which are apparently *without any organic cause*. According to Spence, hysteria phenomena come and go in response of social milieu of the patient, insofar as they are products of social influences on the subject's executive motor system. Indeed hysterical signs appear to be dependent on the patient's ability to attend to its production. In hysteric patients, distraction or sedation reveals the emergence

of normal action, so that the attention is central to the patient's performance of the abnormal act. Spence points out that certain environments encourage the exhibition of hysterical motor signs whereas other environments serve to reverse such behaviors. This means that hysteria is an instance of the 'conspecific' influences on the subject's motor executive system.

In chapter 8, Spence discusses about the cognitive neurobiological basis of an inherently interpersonal behavior: deception. Also deceptive behavior is based on the above examined twofold volitional system. The executive motor system is implicated in producing 'lie' as a novel response and in suppressing the 'true' response by readdressing the value of falsehoods higher than the truths' one. The subordinate slave systems produces the 'true' response, hence it is the 'baseline', the default response of the brain, which is however inhibited by the executive system while deception.

Chapter 9 faces the moral issue whether bad things that human agents do to others are 'chosen' or 'determined'. According to Spence, human beings are not 'perfectible': they are animals who can and will behave 'well' and 'badly', according to their needs and desires: evil and good are both features of human nature. Thereby deviant acts are only examples of abnormal behaviors. Here 'abnormal' has a statistic sense, that is, the characterization of 'normal' depends on our assumptions about what it is that 'most people' do in some specified circumstances. Although 'abnormal' violence may be the result of human response space's decrease determined by contingent factors such as structural/functional anomalies in the perpetrator's brain and genetic abnormalities concerning with neurotransmitter metabolism, Spence recognizes that these biological anomalies are not a sufficient condition for acting badly. The causal power of these anomalies is effective only under specific circumstances (e.g., aberrant influences located within experiential and social spheres), which means that bad behavior is the result of the interaction between the genes, the brain and the social environment.

Finally, in the Epilogue, Spence tries to solve 'Libet's paradox', namely, how can we defend 'free will' if our intentional acts are all *unconsciously* initiated? According to Spence, even though we cannot control our unconscious processes, we can consider an action as morally evaluable if the subject is consciously aware of his/her actions. In conscious awareness, (a) the subjects feel like they are controlling conscious thoughts, and (b) they are conscious of what they are thinking or doing. Consequently: «without consciousness, we cease to be moral agents» (p. 382).

In summary, Spence's book serves as an excellent source book for those philosophers who are interested in neurobiological and naturalistic foundations of free will. His model of human response space may be seen as a characterization of free will in the terms of cognitive neuroscience in order to elaborate a compatibilist view on free will, which attempts to conciliate free will with determinism of natural sciences. Hence his account on free will reminds of Hobbes' compatibilist defense, where free will is not conceived as the subject's capacity of choosing otherwise, but rather of acting *without coercion*, according to his/her own needs and desires.

Book Review

Cognitive Systems and the Extended Mind[*]

Robert D. Rupert
Oxford University Press, Oxford, 2009

Mirko Farina[**]
m.farina@sms.ed.ac.uk

Cognitive Systems and the Extended Mind defends an embedded view of cognition according to which cognitive processes can causally depend on environmental resources in hitherto unexpected ways but cognitive systems are located inside the heads of biological organisms. Rupert firmly rejects the thesis of the extended mind (EMT) according to which cognitive processes can sometimes spread across brain, body and world. In *Cognitive Systems* he surveys most of the arguments that have been offered in defense of EMT and makes a valiant (but ultimately unsuccessful) attempt at refuting them one by one. The volume however isn't entirely negative. Part three of the book develops in detail Rupert's alternative somewhat conservative view of cognitive systems as computational systems located inside the heads of organisms, and of cognitive processes as the manipulation and transformation of internal mental representations.

The volume is organized around three main parts. Part one of the book is mostly methodological and largely devoted to the investigation of how to demarcate cognitive processes from non-cognitive background conditions. In chapter two, in the attempt to delineate cognition, Rupert introduces three desiderata (conservatism, simplicity and scientific feasibility or empirical progress) that are used as theoretical virtues to set up grounds for distinguishing a causal contribution from a cognitive one. Rupert deploys these three desiderata against EMT. Particularly, he asserts that all the criteria EMT enthusiasts have put forward for determining the boundaries of a cognitive system fail to meet his three theoretical virtues. The only account of

[*] I owe an immense debt of gratitude to Julian Kiverstein for his influence on my work. Thanks too to Andy Clark, John Sutton and Erik Myin for their very helpful comments on a previous draft of this review. Any remaining errors are of course down to me.

[**] School of Philosophy, Psychology and Language Sciences, University of Edinburgh

the cognitive, he thinks, that satisfies his theoretical virtues is an embedded one, which takes cognitive systems to be characterized as an integrated set of internal, biological mechanisms interacting with an ever-changing cast of external materials to produce intelligent behavior.

Part two of the book looks at the arguments for EMT and attacks them with a view to defending a system-based approach. More specifically, Rupert calls into question the functionalist credentials of EMT defending the view that the causal roles which are definitive of our mental states should be individuated using the fine-grained details of human psychology as described by cognitive psychology as our benchmark. Chapter five assesses a number of empirical studies that have been used to motivate EMT and shows how these studies can be reinterpreted so as to support an embedded account of cognition. Rupert further emphasizes the virtues of his system-based approach and argues that «the acceptance of the embedded alternative encourages researchers to keep clearly in mind the important asymmetries (between the organism and the external resources), while in no way encouraging them to neglect the interface with or heavy dependence on the environment»(p.107). Rupert takes such asymmetries to count against EMT. He reasons that internal biological processes make a causal contribution to cognition that is very different from anything located in the external environment. These differences (between the biological and the external), he continues, preclude the causal contribution of the external environment from counting as cognitive. However according to some proponents of EMT, it is precisely these differences in causal contribution that motivate EMT. The external makes a causal contribution that is different from but complementary to the contributions of processes taking place inside the head. The complementarity of the inner and outer enables us to cognize in new ways that go beyond what we could achieve on the basis of the bare biological processes taken in isolation. I shall tackle this point in greater detail later in this review. For now, let me add that Sprevak (2009) has also noticed a potential stand-off in Rupert's argument. We can redescribe the empirical evidence for EMT in ways that favors an embedded view but all this shows is that the empirical evidence doesn't decide between extended or embedded. It doesn't support EMT but nor does it favor an embedded view.

Part three develops the system-based approach in much more detail focusing on representation and computation. In this section of the book Rupert aims to develop a positive account according to which cognition is located inside the boundaries of the organism. He does so on the basis of the

arguments from empirical success proposed in chapters two and three. These arguments largely rely on methodological considerations, which favor an organismically bounded perspective that seems to provide the best account of explanation in cognitive psychology. Rupert also attempts to demonstrate how his embedded account can accommodate everything that EMT wants to say without abandoning the traditional computational framework. It should be noted that many proponents of EMT such as Clark, Wheeler and Wilson also accept the computational framework. They argue for what Robert Wilson (1994) has dubbed "wide computationalism". The book finishes up with a chapter on embodied cognition and with some work in cognitive linguistics. Rupert carefully distinguishes embodied from extended approaches. The volume centers around two themes. One is the notion of integration, the second is the need for a mark of the cognitive. In the rest of this review, I intend to attack Rupert's account of integration from a complementarity standpoint (Menary 2007, 2010; Sutton 2010; Rowlands 2010) and cast some doubts upon those views that take the mark of the cognitive to be necessary for cognitive science.

Rupert defines a cognitive system «as an integrated set of stable and persisting mechanisms that contribute distinctively to the production of cognitive science's explananda» (Rupert 2010, p. 344). On his account, a mechanism counts as cognitive and therefore becomes a part of an integrated cognitive system, when it contributes to the production of a wide range of cognitive phenomena across a variety of conditions. The organism is taken as an integrated physical entity whose persistence and relative durability explains the persistent appearance of the integrated set of cognitive capacities realized by the organism itself. But there are other ways of thinking about integration that are consistent with the soft-assembly of cognitive systems on the fly, and that therefore call into question Rupert's persistence and durability requirement.[1] Different components of a soft-assembled system can play quite different roles and have different properties while nevertheless combining to make complementary contributions that enable flexible thinking and acting. A biological cognizer tight coupled with the right kinds of environmental resources, can permit the organism to perform cognitive functions that it wouldn't be able to accomplish in the absence of such external resources. This tight coupling can provide the right kind of temporary integration of the

[1] See Clark 2001, 2003; Sutton 2010; Menary 2007.

internal and external for the organism to accomplish its goals. There is no need for the integration to last beyond the organism's successful performance of a task. The asymmetry point aforementioned becomes particularly relevant here. Extended cognition doesn't require a fine-grained functional isomorphism between inner and outer processes. We get something new by working in mutual partnership with the external environment, something that we wouldn't get from the biological taken on its own. The complementarity between the internal and external therefore «directs our attention to rich, full, and often idiosyncratic cognitive ecologies in which the computational power and expertise is spread across a heterogeneous assembly of brains, bodies, artifacts, and other external structure» (Sutton *et al.* 2010, p. 6).[2] This is why, I think, the human biological cognizer and the environment need to be taken as a complementarily integrated system of cognitive analysis, with neural, bodily and environmental components making equal contributions in the performance of cognitive tasks.

The second central theme of the book is the mark of the cognitive. The need for a mark of the cognitive has been famously promoted by Adams and Aizawa (2008), Rupert (2004) and more recently endorsed by Weiskopf (2010a, 2010b). Its necessity has been postulated, even if with dissimilar goals, by some friends of EMT.[3] However, Clark (2008, 2010a, 2010b) and Sutton (2010) have resisted this claim on the grounds that it unnecessarily complicates EMT. Let me dig a bit more into this. According to Rupert, EMT needs a mark of the cognitive if it is to succeed in arguing that the environment is playing a constitutive role in the emergence of a cognitive process. The need for such a mark follows from the necessity to distinguish factors that are genuinely parts of a cognitive system from factors that only causally contribute and don't have any constitutive involvement. Rupert individuates the locus of such a mark in the organism and his view of cognitive systems, discussed above, is taken as a measure to distinguish what is cognitive from what is not. Particularly, Rupert believes that what happens within the biological cognizer (the set of mutual interrelations between body and brain) can entirely account for cognition. What is external to the bio-physical architecture of the organism can only ever make a causal contribution. The debate around the mark of the cognitive famously emerged from the discussion of the causal-constitutional

[2] Also see Hutchins 2010, Tribble and Keene 2010 for similar arguments.
[3] See Rowlands 2008, and particularly Wheeler 2005 and forthcoming.

conflation. Some people assume (and Rupert certainly stands among them) that this conflation entails the need for a mark of the cognitive. Rupert provides one that entails that an external resource can make a cognitive contribution to behavior only if it corresponds in a fine-grained way with the causal contribution of our inner states. There are specific psychological effects for instance (e.g., negative transfer, primacy and chunking effects) that we do not find in cases of extended memory. Because of this failure of fine-grained correspondence we shouldn't treat use of external resources in memory tasks as cognitive uses. We should say instead that the external resource is only making a causal contribution.

Now, as I stated above, Clark (2008, 2010a, 2010b) believes that the attempt to identify a mark of the cognitive is unlikely to bear fruit.[4] First of all Clark denies that the differences in fine-grained functional role of the external and internal matter, arguing instead that the sort of functional equivalence that counts for the parity argument is determined at a fairly coarse-grained level. If the cognitive were marked out by a fine-grained correspondence, this would prevent us from attributing cognition to creatures that are appreciably different (either biologically or psychologically) from us. The demand for a fine-grained correspondence requires us to scale new heights of neurocentrism and anthropocentrism. Cognition, as far as Clark is concerned, does not necessarily necessitate minds that work in the same fine grained ways as human minds work. Additionally, since the differences between «external-looping (putatively cognitive) processes and purely inner ones will be *no greater than those between the inner ones themselves*» (Clark 2010a, p. 51), it is likely that the inner goings-on, postulated by opponents of EMT, will turn out to be a motley crew. Clark has in fact brilliantly noticed, that we already possess a practical grasp on the kinds of coarse-grained behavior patterns that we presume to be characteristic of key cognitive processes, such as the holding of a standing (dispositional) belief (Clark 2010b). A very basic and relatively liberal appeal to folk psychology would therefore suffice to guide us in working out what counts as cognitive and what does not. Wheeler has recently disagreed arguing that EMT needs «a scientifically informed, theory-loaded, locationally uncommitted account of the cognitive» (Wheeler 2010). Clark has responded that such a quest is unnecessary and unlikely to succeed. The shape and the contour of any such a theory will always and ultimately be determined

[4] Also see Sutton 2010 and Menary 2007 for similar arguments.

by what one takes as central examples of real-world realizers of cognitive processes (Clark 2010b).

To conclude: *Cognitive System and the Extended Mind* raises some significant challenges for EMT and provides powerful support for a more traditional orthodox approach to cognitive science. The book is not for everyone: it is densely written and some of its arguments remained cryptic at least to this reader. It nevertheless succeeds in making a strong case for an embedded perspective even if Rupert's opponents are unlikely to be convinced. Rupert's arguments remain undecisive. He fails to point to any way-out from the impasse in which the debate between embedded and extended has fallen that favors Rupert's embedded conservatism over the more radical ideas of friends of the extended mind.

REFERENCES

Adams, F., & Aizawa, K. (2009). *The Bounds of Cognition.* Oxford: Wiley Blackwell.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science.* Oxford: Oxford University Press.

Clark, A. (2003). *Natural Born Cyborgs, Mind, Technologies and the Future of Human Intelligence.* Oxford: Oxford University Press.

Clark, A. (2008) *Supersizing the Mind: Embodiment, Action and Cognitive Extension.* Oxford: Oxford University Press.

Clark, A.(2010a). *Memento*'s revenge: The extended mind, extended. In R. Menary (Ed.), *The Extended Mind*, (pp. 43-66). Cambridge, MA: MIT Press.

Clark, A. (2010b). Finding the mind. *International Journal of Philosophical Studies*, 1-15.

Hutchins, E. (2010). Cognitive ecology. *Topics in Cognitive Science, 2*(4), 705-715.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded.* New York: Palgrave Macmillian.

Menary, R. (2010). *The Extended Mind.* Cambridge, MA: MIT Press.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology, 22*(1), 1-19. [2008]

Rowlands, M. (2010). *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *The Journal of Philosophy, 101*(8), 389-428.

Rupert, R. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research, 11*(4), 343-356.

Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy, 106*(9), 503-527.

Sutton, J. (2010). Exograms and interdisciplinarity: history, the extended mind, and the civilizing process. In R. Menary (Ed)., *The Extended Mind*, (pp.189-225). Cambridge, MA: MIT Press.

Sutton, J, Harris, C., Keil, P., & Barnier, A. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenom Cogn Sci, 9*(4), 521-560.

Tribble, E., & Keene, N. (2010). *Cognitive Ecologies and the History of Remembering: Religion and Education in Early Modern England*. London: Palgrave.

Weiskopf, D. (2010a). The Goldilocks problem and extended cognition. *Cognitive Systems Research, 11*(4), 313-323.

Weiskopf, D. (2010b). Review of *Cognitive Integration: Mind and Cognition Unbounded*. *Mind, 119*(474), 515-519.

Wheeler, M.(2005). *Reconstructing the Cognitive World*. Cambridge, MA: MIT Press.

Wheeler, M. (2010). In defence of extended functionalism. In R. Menary (Ed.), *The Extended Mind*, (pp. 245-270). Cambridge, MA: MIT Press.

Wheeler, M. (forthcoming). *Extended X: Recarving the Biological and Cognitive Joints of Nature*. Cambridge, MA: MIT Press.

Wheeler, M., & Clark, A. (1999). Genic representation: Reconciling content and causal complexity. *Br J Philos Sci, 50*(1), 103-135.

Wilson, R. (1994). Wide Computationalism. *Mind, 103*(411), 351-72.

Book Review

Embodied Cognition

Laurence Shapiro
Routledge, London, 2010

Andrea Danielli *
andrea.danielli@collegiodimilano.it

1. The Book and Its Merits

Shapiro has the extraordinary merit of analyzing a highly debated subject, as embodied cognition, with honesty and cold blood. He is able to disentangle fascinating views comparing all the supportive arguments and experimental evidences without prejudices.

The structure of the book is very clear and effective in organizing the large literature with all its trends; embodiment is decomposed over three themes: conceptualization, replacement, and constitution. *Conceptualization* gives accordance to the idea that «the properties of an organism's body limit or constrain the concepts an organism can acquire» (p. 4). The concept of *replacement* bases on this claim: «an organism's body in interaction with its environment replaces the need for representational processes thought to have been at the core of cognition» (p. 4), finally *constitution* assert that «the body or world plays a constitutive rather than merely causal role in cognitive processing».

Of course it is a choice with some arbitrariness, as in the case of including system dynamics in the replacement theme. Certainly dynamicists discard representations and insist on the coupling between brain, body and the environment, but we see several exception as Van Gelder and Port's moderate claim «a wide variety of aspects of dynamical models can be regarded as having a representational status: these include states, *attractors*, trajectories, bifurcations, and parameter settings» (Van Gelder and Port 1995, *my emphasis*), as well as Edelman and Izhikevich (2008) that analyze in their

* University of Paris 1 Panthéon-Sorbonne

model only brain dynamics, with no regard to environment. Another reason to be cautious resides in the great novelty of system dynamics: their conceptual and mathematical tools. These tools are not extendable to all the embodiment paradigm: in terms of set theory, between embodiment and system dynamics we have an intersection, not an inclusion.

2. Some (Little) Criticism

I found disputable most part of the chapter devoted to cognitive sciences: it is too brief, only twenty pages, it holds on old case histories, with no historical treatment at all. Some remarks about cognitive sciences' origins: without a brief account of behaviorism it is difficult to understand the novelty, and I would had like just a few words about functionalism. For what it concerns case histories, why is Shapiro talking only about Newell and Simon 1961's research, when we progressed through fuzzy logic, heuristics, intelligent agents, data mining? In this strange arbitrariness, Shapiro did not talk of object recognition (Marr, Tarr, Biederman), nor language acquisition (generative grammar is the best didactical example to explain cognitive sciences).

Finally, Shapiro should had invested more time to talk about representations, moving from classical treatment and penetrating the neuroscientific approach as did Bechtel (2008).

3. Simulation

When talking about *conceptualization*, Shapiro admits that bodily characteristics may well be simulated by an algorithm, and this induces him to conclude: «embodiment is not inconsistent with computationalism» (p. 93). Unluckily, when examining the *envatment argument*, which opens the possibility to generalize this statement, Shapiro reduces its range to a lesser extent.

That is a pity: simulation is a clear concern for embodied cognition, because it shifts attention from the body to the brain, where information is really processed. As in phantom limb syndrome, what counts is not the origin of information (that may not exist), but its elaboration. I believe that simulating the brain itself discloses the opportunity of a computationalism without representations, indeed dynamicists use software to model brain at neural level. These are the early steps to make neural mechanisms' simulation a means

to shed light on information processing.

4. Competing Paradigms

Is embodied cognition a unified body of knowledge, a new promising paradigm? I do not think so, and I agree with Shapiro conclusions, as he finds conceptualization and replacement loosing the challenge with cognitive sciences. In fact, the real changes will arrive from neurosciences, especially when we will be able to correctly read brain coding, using fluorescent optical imaging, in vivo single cell recording, or new nanotechnological techniques still at conceptual development. New data will oblige us to change definitely our old assumptions. I am not proposing to quite our theoretical attitude, waiting for brute powerful solutions, but I do not share this author's optimistic claim «work on Conceptualization is ongoing, and neuroscientific findings promise to energize some of its basic assumptions» (p. 210). The use of neuroscientific knowledge seems an improper attempt to revitalize a gloomy paradigm. Instead of this risky strategy, philosophers shall focus on epistemology of the cognitive sciences, dissecting methods and conceptual tools.

A simple comparison between different paradigms, both diachronic and synchronic, allows us to see at the same time the inadequacy and fragmentation of the cognitive domains of research. Indeed, every cognitive paradigm works well over few cognitive capacities: computationalism started with problem solving, connectionism is good at describing learning mechanism from complex patterns, embodiment is perfect at explaining action and motion control. When we try to extend those paradigms beyond the border they fail completely their explicatory mission. I would have liked if Shapiro had developed this statement further to account for this fiasco: «I think that an effort to cover all the evidence under a single umbrella is not likely to succeed» (p. 2).

References

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum Associates.

Izhikevich, B. E., & Edelman, G. M. (2008). *Large-scale model of mammalian thalamocortical systems. PNAS, 105*(9), 3593-3598.

Van Gelder, T., & Port, R. F. (1995). *Mind as Motion.* Cambridge, MA: MIT Press.

Book Review

Siamo davvero liberi?
Le neuroscienze e il mistero del libero arbitrio

M. De Caro, A.Lavazza and G. Sartori (Eds.)
Codice Edizioni, Torino, 2010

Giuseppe Vicari *
giuseppevicari@unipa.it

The book focuses on the impact of neuroscience on our conception of free will as the capacity of an agent of rational self-determination given a set of possible alternative courses of action (pp. IX, 111): a theme which is a core part of the more general issue whether the central traits of human beings as biologically embodied and socially embedded mindful, rational, morally responsible agents can be adequately located inside a naturalistic conceptual framework, or if the development of neuroscience will force us to get rid of them.

The two parts of the work – "The fall of the ancient certainties" and "Theoretical horizons and social perspectives" – review the relevant results of scientific investigations and discuss their theoretical and social implications, together with a critical discussion of the interpretation of the empirical evidence.

The well-known starting point is given by Libet's studies of the neural mechanisms underlying the production of voluntary movements, where subjects are required to perform a simple wrist movement whenever they feel the desire to do so and to report the onset of the conscious decision by indicating the position of a dot revolving on a clock located in front of them.

This setting allowed Libet to compare the onset of conscious decision with:

a) the onset of the "readiness potential", the increased electric activity in the supplementary premotor cortex which precedes voluntary movements and is detected through EEG; and

* University of Palermo

b) the onset of the movement, identified through electromyography.

The readiness potential starts 350 ms before the conscious decision takes place. It seems, then, that the conscious decision is causally inert with respect to the performance of the movement. In fact, so the reasoning goes, the conscious decision is made only after the onset of the relevant (neural) causal chain: therefore, in short, the decision has not been made by the subject, but rather by "his/her" brain.

As John-Dylan Haynes points out this methodology can be criticized with regard to both the reliability of subjective reports as measures of the onset of the conscious decision[1] and with respect to the assumption that the readiness potential is a causally sufficient antecedent of the action. The latter point, in particular, is due to the fact that Libet's studies focused only on the supplementary motor area. Moreover, the experiment tests only one of the variables involved in conscious decision making – the "when", but not the "what" of the action.

The subjects of Haynes' study choose whether to push the left or right button whenever they want while looking at a series of letters appearing on the screen located in front of them. When the task is performed the subject chooses the letter seen while he/she was making up his/her mind. In this way the onset of the conscious decision can be determined without relying on the possible distortions of subjective timing. The finding of the study is that the frontopolar cortex contains information predictive of the content of the decision from seven to ten seconds before the decision is made, with an accuracy of 60%. It seems, then, that a set of unconscious neurobiological processes takes place over time before the conscious decision is made, and that it contributes to prepare decisions experienced by the subject as free and as occurring in a single moment in time (p. 16).

What is the role played by conscious intentions in this framework? According to Daniel Wegner folk psychology cannot account for the complexity of behavior as it is revealed in borderline cases such as automatisms (like anarchic-hand syndrome, where apparently goal-directed movements are produced against the subject's conscious will) and delusions of control, where subjects think of themselves as authors of actions that they have not, in fact,

[1] See Filippo Tempia's contribution for an excellent discussion of this point, especially pp. 92-99.

done.[2] These cases suggest that thought and action could result from distinct mechanisms (pp. 26ff.), where the ones responsible for behavior would drive online performance of actions, while the thought-producing mechanisms would just give the agent a first-person prediction, or indication, of what is going to happen (pp. 40-41). From a causal point of view the will is nothing but a useful illusion, an "emotion of authority" (pp. 46-48) allowing the organism to distinguish between what he/she is doing from what other organisms are doing in a shared physical and social environment.

Mental causation is, then, "narrative" or "apparent" (p. 39): according to the studies reviewed by Rigoni and Brass (pp. 73-75) a conscious intention of acting would be just a process of inferential reconstruction partially based on events taking place after the performance of the action. In a typical experiment (p. 73), TMS application over the presupplementary motor area (PRE-SMA) 200 ms after the action causes the subject's perception of the onset of the intention to shift backward in time, while the perceived performance of action shifts forward in time. This would prove that PRE-SMA activity taking place after the action is relevant for the perception of the intention and, therefore, for its "construction".

The second part of the work starts with Filippo Tempia and Roberta De Monticelli's different but convergent criticisms of the standard interpretation of the studies illustrated in the first part.

Tempia points out that the standard interpretation would rely on a dualistic model of the relationship between conscious and neurobiological processes. This model would postulate, dualistically, separate and mutually inconsistent mental and physical causes where the former, if free will and voluntary actions have to be genuine phenomena, are supposed to occur before and independently of the latter.

Given this basic dualistic framework, it is obvious that Libet's results can be interpreted as a "scandal": saying that the decision-maker is the brain and not "you" makes sense only if one assumes that the conscious will is a sort of *causa sui* separated and independent from brain processes (pp. 88-90, 100ff.).

Tempia opposes this model to the one exemplified, in physics, by magnetic fields, where cause and effect are simultaneously realized: an electric discharge

[2] Cf. Wegner and Wheatley 1999.

creates the field which in turn affects the discharge, with cause and effect simultaneously realized.[3]

According to this model the neural processes examined by Libet, Haynes and others do not correlate with conscious will, but rather with other stages of voluntary action preparation, such as recalling to memory the instructions of the task or translating these instructions into a motor performance (p. 101). Moreover, this model would take into account the scientific evidence showing that emotions play a direct causal role with respect to the modulation of practical rationality and social interactions (pp. 102ff).[4]

Of course, that the ontology of mind underlying Libet's studies is controversial, and probably contradictory, has already been noted by Dennett (2003)[5] and by those contemporary philosophers of mind interested in the problem of mental causation, from Donald Davidson to Jaegwon Kim's dilemma of causal exclusion.

However, although these criticisms usually lead to materialism as the right solution of the mind-body problem, we could observe – following the spirit and, I believe, the letter of Tempia's proposal – that inferring epiphenomenalism from a neural explanation of behavior reveals an implicit *a priori* acceptance of dualistic categories, such that it would be impossible that consciousness itself is a higher-level brain process. We might say, as Searle once put it, that materialism (the denial of any ontological and/or causal reality to the mind as such) is in this sense «the finest flower of dualism» (Searle 1992, p. 26).[6]

[3] Similarly, John Searle has argued that the mechanism of bottom-up, no time gap causation exemplified in physics and biology gives us a general theoretical model of the ontology of mind (which he dubs "biological naturalism") which takes into account both the ontological irreducibility of the subjectivity of mind and its causal reducibility. For a systematic analysis of this model even with respect to the problem of mental causation, see Vicari 2008.

[4] Cf. Damasio 1994 and 1999.

[5] Cf. De Monticelli's contribution, p. 106.

[6] More recently Searle (2001, pp. 288-289) has integrated Roger Sperry's model of top-down causation within his "biological naturalism". He also points out, however, that the uncritical use of metaphors such as "bottom-up" and "top-down" causation could be misleading in this context because it suggests the existence of mutually separated and independent "mental" and "physical" causal chains and obscure the fact that consciousness is a "system feature" of the brain (Searle 2001, p. 287). As such, consciousness can affect the behavior of the elements of which the system is composed without postulating any breakdown in the causal closure of physics. For an analysis of the notion of "systemic causation", see Di Lorenzo Ajello 2009.

According to Andrea Lavazza and Luca Sammicheli a dualistic model of mind would implicitly shape our legal systems. Concepts such as "being chargeable", "being responsible", "being guilty" are grounded on the assumption of free will. Neuropsychological tests are already used in criminal trials to establish the ability of a person to understand and will under the presupposition that free will exists, though it might break down.[7] But what happens if neuroscience shows that "the real decision-maker" is just a set of "material causes" such as automatic neural mechanisms? Again, if the juxtaposition of "I" and "my brain" holds, then the notion of free will is in deep problems (p. 153). It is unclear, for example, whether we would be entitled to hold a retributive view of legal punishment or if we should see crimes as the result of a "malfunctioning" of a system and, then, think of the guilty person simply as a damaged element that must be kept away from society.

While Tempia puts forward an ontologically and scientifically motivated criticism of Libet's results, Roberta De Monticelli works out a phenomenological criticism.[8]

The materialist argues that every event has a causally sufficient antecedent that determines it, and since actions are events, then actions are determined. But this argument leaves out the fact that actions are not experienced as determined events, but rather as *motivated* acts. This experience, she argues, reflects an ontological difference between, for example, falling asleep and going to sleep: the former case is determined by a causally sufficient antecedent, but the latter case – the action case – requires that the agent takes a position toward his/her *motives* and makes them effective through his/her decision (p. 121). But then, if a decision is the act through which a person makes his/her motives effective through an exercise of his/her "positionality", it seems that an unmotivated decision, like the ones typical of the experimental settings, is not a decision at all (pp. 124ff.).[9]

Mario De Caro argues against the re-actualization of emotivism as a reductive explanation of morality put forward by Chapman and colleagues (Chapman *et al.* 2009), who claim that "moral disgust" is strongly associated

[7] Cf. Gnoato and Sartori's contribution.

[8] For a criticism of Wegner based on his misleading description of – or lack of attention to – the phenomenology of agency, see Bayne 2006.

[9] Cf. Tempia's contribution, pp. 97-99.

with the evolutionarily more primitive emotional reaction of disgust that one has, for example, while standing in front of a rotten food.

The claim is plausible provided that one does not interpret it – as Chapman and colleagues apparently do – as the claim that the evolutionary account of the mechanisms enabling morality offers *ipso facto* an explanation of its contents. This is a stronger thesis, even because it implies the reduction of the normative concepts of morality to the non-normative concepts of emotional reactions and evolutionary neuropsychology.

For example, if moral judgments were reducible to emotional reactions we could not understand why two persons, endowed with different cultural systems, could react in quite different ways (like moral disgust and moral approval) to the same event and context: in cases like these the order of explanation could plausibly go from culture to physiology, while it would be useless to explain the different moral reactions in terms of emotional reactions because the latter should in turn be explained by something else capable of taking into account the normative character of moral judgments (pp. 137-139).

Morality requires the possibility of a detached look at our own emotional and instinctive reactions because without it we would not be able to understand that a given reaction is not only wrong with respect to a certain context, but also, in a deeper sense, morally, normatively, rationally unacceptable. And the deep reason of this fact is that, as De Caro also argues, morality requires the articulation of reasons justifying a moral judgment in the language game of asking and giving reasons, while nothing similar seem to be required to emotional disgust – especially when food-related (pp. 142-145).

De Caro's argument, at least as I understand it, does not deny, for example, the significance of Damasio's analysis of Phyneas Gage's case as showing that emotions play an active causal role in shaping the rationality of our social interactions. Rather it points out the difference between a condition enabling some capacity to work, and the reductive identification of the capacity itself with its causal precondition.

Being characterized by a multidisciplinary approach, this book offers the reader a highly detailed while accessible picture of the problems, of the different theoretical views, of the arguments supporting the views and of their implications for our self-conception. As such it provides the reader with a useful tool to find one's way in an extremely stimulating, rich and complex debate. Whether we are free remains an empirically and conceptually open

question, as some of the essays here collected convincingly argue, and, as Adina Roskies argue in her contribution, perhaps neuroscience will just explain the mystery without explaining the phenomenon away.

REFERENCES

Bayne, T. (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. In S. Pockett, W. Banks & S. Gallagher (Eds.), *Does Consciousness Cause Behavior?* (pp. 169-185). Cambridge, MA: MIT Press.

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In Bad Taste: Evidence for the Oral Origins of Moral Disgusts. *Science, 323*(5918), 1222-1226.

Damasio, A.(1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Gosset/Putnam Press.

Damasio, A. (1999). *The Feeling of What Happens*. San Diego, CA: Harcourt.

Dennett, D. (2003). *Freedom Evolves*. New York: Viking.

Di Lorenzo Ajello, F. (2009). La mente in azione. Introduction to J. R. Searle (U. Perone, Ed.), *Coscienza, linguaggio e società*, (pp. 9-17). Torino: Rosenberg & Sellier.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Searle, J. R. (2001). *Rationality in Action*. Cambridge, MA: MIT Press.

Vicari, G. (2008). *Beyond Conceptual Dualism. Ontology of Consciousness, mental Causation, and Holism in John R. Searle's Philosophy of Mind*. Amsterdam/New York: Rodopi.

Wegner, D. M., & Wheatley, T. P. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54*(7), 480-492.

Book Review

Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness

Alva Noë
Hill and Young, New York, 2009

Marco Spina *
mar.spina@gmail.com

As the author notes at the beginning of the book,

> we live in a time of growing enthusiasm for the brain. Perception, memory, pleasure or displeasure, intelligence, morality [...] the brain is supposed to be the organ responsible for all of it. It is common belief that even consciousness, the Holy Grail of philosophy and science, will soon become the object of a neural explanation. (p. XI)

As Patricia Churchland, one of the most important experts of philosophy of neuroscience, has noted: «the weight of evidence now implies that it is the brain, rather than some non-physical stuff, that feels, thinks, decides» (Churchland 2002, p. 1).

However, after decades of common endeavors on the part of neuroscientists, psychologists and philosophers, the only point that seems non controversial about the role of the brain in making us consciousness – that is, the way it brings upon sensations, feelings, and subjectivity – is that we know nothing about it. This is what the American philosopher Alva Noë states in his book. In this book the author deals with the problem of consciousness, suggesting a radical solution: abandoning the assumption that, ever since Descartes, confines the mind within the brain.

Thus, the idea that the only properly scientific inquiry of consciousness is the one that identifies it in the events occurring within the nervous system collapses. Accordingly, Noë suggests a new thesis consisting in the claim that, in order to understand consciousness – meaning the fact that we think, feel and

* Paris-Sorbonne University (Paris IV)

that a world manifests itself in front of us – we need to look at a larger system of which the brain is only element. Consciousness is not something the brain achieves on its own. Consciousness requires the joint operation of brain, body and world.

Two notable aspects surface from these lines: the first one concerns what has been sometimes called the "gap" or "explanatory blank" (which scientists haven't yet been able to fill) between neural states and conscious experience. The central thesis of this book is to recognize that the brain, in itself, is not a source of experience or cognition. Experience and cognition are not bodily by-products. The second aspect sheds light on the new task of philosophy: in order to advance in the understanding of consciousness one must abandon the internal neural micro-analyses and look instead at how each of us, in his wholeness, carries forth his life in the surrounding world, with it and in response to it.

This is why Alva Noë tries to show the assumption that the current research in neuroscience is badly mistaken. Indeed, as the first chapter of the book shows, the brain (on its own) doesn't explain what we are:

> my consciousness now – with all its particular quality for me now – depends not only on what is happening in my brain but also on my history and my current position in and interaction with the wider world. (p. 4)

In light of these considerations, the author assumes a critical attitude towards all those theories, scientific or philosophical, that deal with "the problem of other minds" (the way we decide who or what is conscious) starting from a theoretical point of view. The attack is directed especially toward the theory of mind, accused of treating consciences like something private and hidden. There is an experiential and essentially practical reason for our faith in the existence of the minds of others. No "sane" person could take seriously the idea that our knowledge of other minds is merely hypothetical. However weak our proof that others possess a mind may be, it would be altogether absurd to think that because of this our commitment to the existence of others' minds fails.

A simple example is enough: Noë notes, in this regard, that the relationship between the young child and whoever takes care of him is truly paradigmatic. There is no theoretical distance between the child and the caregiver. The child does not question whether his mother is an animated being or not. For the child the living consciousness of his mother is simply

something present, like her warmth or her breath; it is, in part, what animates their relationship. The mind of the mother and that of the child manifest each other in the direction that is made up of tenderness and cuddling. This is why, if one wants to speaks of a commitment to the alive consciousness of others here, one should speak not of a cognitive commitment but, rather, of a practical commitment.

This statement sheds light on two very important aspects of Noë's provocative and stimulating theoretical suggestion: the first one is that like the child in his relationship to the mother, we are involved in one another; it's our living together that assures our living conscience of other people. The second consideration regards the clarification of the field of enquiry within which the philosopher's speculation occurs. Indeed, Noë shows that there is something paradoxical in the science of the mind: scientific knowledge requires a detached attitude, but the mind become object of study only if we assume a different attitude, that is, a much more involved one.

The example of the relationship between mother and young child illustrates well that the perspective that we need, from which the meaningful, non mechanical nature of conscious life can come into focus, is none other than the biological perspective. To understand an organism, we must take up a perspective on its life that is at once narrative and historical and also ecological. This shows that the question of consciousness arises for living beings and it arises for them because living beings exhibit at least primitive agency. To study mind, as with life itself, we need to keep the whole organism in its natural environmental setting if focus. Neuroscience matters, as does chemistry and physics. But from these lower-level or internal perspectives, our subject matter loses resolutions for us.

The aim of this book is therefore to convince the reader that there is something perverse in believing that we are our brain, that the world we have experience of is within us. We don't need to have the world inside us: we have access to the world that surrounds us; we are open to it. The idea according to which we are our brain is not something scientists have learnt; instead, it is a prejudice that they take from home into their laboratories.

Once any type of prejudice has collapsed, at the end of the book, Noë outlines a new conception of conscience: the substratum of our life and our experience is nothing but the world we live in. The entire world and the nature of our situation within it are the material of a theory of conscious life. In this story, the brain has the leading role, no doubt. But the task of the brain is not to

"generate" consciousness. Consciousness is not this type of thing. Consciousness is not a thing. This is what a genuine biological approach to the study of human mind and human nature teaches us. So if we want to understand consciousness we must turn our backs on the orthodox conception according to which consciousness is something that happens within us and we must make progress in the creation of an authentic ecological theory of ourselves. This is why, it is now clear that consciousness is achieved in action, by us, thanks to our situation in and access to a world we know around us.

Thus this book – rich of reflections and argumentations, lucid and systematic in examining (and confuting) the different positions on the subject – don't fail to make readers reflect and to provoke discussions really helping to understand in which way the encounter between our brain and our experiences allow us to become the people we are.

REFERENCES

Churchland, P. S. (2002). *Brain-Wise: Studies in Neurophilosophy*. Cambridge, MA: MIT Press.

Book Review

Radical Embodied Cognitive Science

Anthony Chemero
MIT Press, Cambridge (MA), 2009

Silvano Zipoli Caiani *
silvano.zipoli@unimi.it

Three tenets make Chemero's embodied cognitive science "radical", as stated in the title: *anti-representationalism*, *direct perception* and *realism*. The combination of these three assumptions offers something excitingly new to several fields of philosophy such as phenomenology, theory of perception and epistemology of cognitive science. This book also represents an intriguing challenge to many established ideas in philosophy of mind, especially representationalism and computationalism. Indeed, Chemero's book is an ambitious work, aiming to become in the field of embodied psychology what Fodor's famous book *The Language of Thought* was for computational psychology. Accordingly, a great deal of the book is devoted to dismantling the Fodorian paradigm.

The book is divided into four sections. The first section (chapters 1-2) introduces the author's dissatisfaction with traditional arguments in cognitive science. In the second section (chapters 3-5), Chemero presents an alternative to representationalism in philosophy of mind, namely, a *dynamical approach* to cognition. In section three (chapters 6-7), the author attempts to define ecological psychology as the background theory for his Radical Embodied Cognitive Science (hereafter, RECS). Finally, section four (chapters 8-9) investigates some philosophical consequences of reductionism and realism in cognitive science. Let me now introduce and comment on each chapter individually.

Chapter One is a nice introduction to what Chemero calls the "Hegelian arguments", that is, arguments based on theoretical posits and no empirical

* University of Milan

evidence, stating (a priori) that some particular explanatory approach will certainly fail. According to Chemero, there are currently numerous Hegelian arguments in the field of cognitive science; it is therefore a field characterized by several contrasting theoretical frameworks, each aspiring to establish itself as the main research paradigm, even if it lacks any empirical support. Although Chemero contrasts such a priori approaches in cognitive sciences, he encourages theoretical pluralism as a positive condition for the development of a scientific discipline. Referring to Feyerabend's epistemological analysis, Chemero emphasizes the fact that the presence of many competitors enhances scientists' ability to deal with the empirical findings of rivals, providing new potential falsifiers and more refined interpretations.

The second chapter proposes a taxonomy that enables Chemero to frame his conception within the contemporary debate in philosophy of mind. The author thus distinguishes between *representationalist* and *eliminativist* approaches to the mind. The former is characterized by the assumption that there are mental representations that stand for external things in the world. The latter, on the contrary, assumes that cognition does not mirror the world and should be understood as a vital function of the animal. Based on this distinction, Chemero's conception emerges as the result of an *eliminativist* choice where the agent and the environment are intended as two coupled systems that cannot be modeled as a set of separate parts. Following this line, Chemero considers the body and the environment as a *dynamical system* constituted by variables that change according to mathematical laws. This makes it possible to account for cognitive processes through differential equations that pair animal parameters with environmental parameters. Here, Chemero also introduces Randy Beers' model of the artificial agent and the Van Rooij *et al.* account of imagined actions as two examples of how dynamical systems have the power to explain different cognitive situations without relying on the concept of mental representations.

In chapter three Chemero considers different accounts of representation, presuming Millikan's conception is representative of the entire field. This functions as an introduction to Chemero's argument against representationalism. It should be noted that Millikan account of representation is characterized by a teleological approach, which means it can be classified as a non-radical theory of representation. As such, what Chemero faces is a definition of representationalism where the question is not simply whether a neuron, or a portion of the nervous system, codifies for something in the

world; he aims to show the supremacy of a non-representational approach even over non-naïve theories of mental representation.

In chapter four Chemero introduces his argument against representationalism in cognitive science. He initially distinguishes between two different anti-representationalist stances: the metaphysical and the epistemological. The metaphysical claim is that nothing in the nature of a cognitive system is a representation; the epistemological stance, on the other hand, is that we need not resort to mental representations in order to explain cognitive processes, without assuming anything about the nature of the cognitive systems itself. Chemero's point is that endorsing a metaphysical stance doesn't add relevant information to a dynamical account of a cognitive process. According to Chemero, even if representational accounts of cognitive systems are possible, that is, even if a cognitive process may be interpreted as positing a role for representations, they appear superfluous and unnecessary when a dynamical account is also available. It is effectively an argument of simplicity (like Ockham′s razor), where dynamical descriptions are considered simpler, more complete accounts of cognitive processes, while mental representations are considered nothing but redundant entities. According to this view, a representationalist approach to cognitive systems is superfluous only when a complete dynamical description has actually been developed, so as Chemero himself notes, how much of cognition can be accounted without reference to "representational glosses" is a matter of fact.

According to this empirical characterization, a potential problem for a dynamical account arises. Given the Humean roots of dynamical cognitive science, according to which no unobservable entity should have an explanatory role, one could argue that it doesn't provide a useful guide to predictions and new discoveries. In order to face this problem, chapter five is dedicated to a defense of the heuristic value of anti-representationalism in cognitive science. With this purpose in mind, Chemero analyzes the Haken-Kelso-Bunz dynamical model, showing how this framework is able to produce predictive systems without any reference to mental representations.

In order to make sense of Gibson's ecological psychology as a theoretical background for a dynamical and anti-representationalist approach to cognitive science, chapter six is devoted to introduce the critical notion of *direct perception.* In the first part of this section Chemero explicitly acknowledges his debt to the Turvey-Shaw-Mace approach, which has introduced a new order in the field of ecological psychology. He thus outlines a philosophical account

of Gibson's ecological theory of perception, according to which environment, information and perception determine one another. In the second part of the chapter, Chemero tries to overcome the limits of the Turvey-Shaw-Mace approach (concerning its generalizability and its application to social information) focusing on the unmediated character of perceptual processes.

The assumption of perception as a direct and unmediated process leads Chemero to emphasize the animal's ability to use environmental information to guide actions without necessarily needing mental representations. Drawing from this view, chapter seven focuses on a renewed definition of Gibson's famous notion of affordance, aiming to make it more clear and sound. Chemero endorses a notion of affordance that is actually deeply different from Gibsonian and post-Gibsonian definitions. According to Chemero, affordances are relationships between the perceiver and the environment and cannot be reduced to mere properties of the perceived things. More precisely, Chemero emphasizes the causal role of the perceiver's motor abilities, arguing that the agent's motor repertoire may cause changes in the layout of the available affordances and that the perception of affordances may changes the way motor activities are exercised. Accordingly, perception and action cannot be considered two independent cognitive modules. Rather, perception emerges as a type of action; furthermore, a great deal of action can be considered functional to realize perceptive purposes.

Chapter eight turns to the implications of anti-representationalism for reductionism. Radical reductionism (i.e., physicalism) ignores the ecological character of perception, confining the entire account of cognition to the nervous system. Chemero's RECS focuses primarily on the relationships between action, perception and environmental information, resisting the "brain obsession" that frequently inspires reductionism in philosophy of mind. Chemero also includes in this chapter an analysis of animal exploration based on a comprehensive review of many published papers on this subject, showing that the literature often ignores the ecological character of the object employed in the experiments, and therefore fails to notice their effects on the animal's exploratory behaviour.

Finally, in the last chapter of the book, Chemero defends a realist approach to radical embodied cognitive science. Here, as the author himself notes, the source of the problem is represented by the notion of affordance and its dependence upon the perceiver. The question is: can an affordance be considered an autonomous thing, distinct from the basic furniture of the

world? According to Chemero, affordances are not something pertaining to the domain of subjectivity, nor are they mere properties of external reality. Affordances are relations between the agent's motor abilities and the features of the environment (chapter 7). Therefore, their ontological status appears controversial in light of traditional views such as physicalism or idealism. Referencing to Hawking's entity realism, Chemero argues that affordances are genuine theoretical entities that acquire their value of reality in light of their role in actual experimental practice. This constitutes what can be considered a *pragmatic* stance about scientific realism that makes it possible to disentangle affordance perception from the domain of subjectivity, without committing RECS to an untenable metaphysical notion of reality.

Let me conclude this review with some brief remarks about Chemero's book. The work is certainly a provocative presentation of an alternative to the mainstream representationalism in cognitive science. It provides both an introductive and a "technical" approach to what cognitive science might look like without reference to inner mental representations and computations. Accordingly, Chemero's book is accessible to readers with different backgrounds and from different areas of expertise. Philosophers such as phenomenologists and epistemologists will find many intriguing suggestions concerning the development of a theory of perceptive experience linking traditional pragmatism, ecological psychology and contemporary enactivism. At the same time, scientists confident with questions involving the modeling of perception will find this book an incisive attempt to establish a new framework in cognitive science. The many experimental examples contained in the book represent a challenge to scholars who are still skeptical about a cognitive science that affords no role for mental representations.

As for RECS potentially becoming a mainstream framework in cognitive science in the near future, that depends on the empirical adequacy of its theoretical model. As Chemero himself recognizes in his endorsement of a pluralistic stance in epistemology, RECS is not the sole true account of the mind. Yet it is certainly the most comprehensive conception that links the mind to the body and the ecological order.

Commentary

From the Puzzle of Qualia to the Problem of Sensation

Phenomenology of Perception
Maurice Merleau-Ponty

Roberta Lanfredini *
lanfredini@unifi.it

Phenomenology of Perception is the expression of Merleau-Ponty's epistemological and methodological perspective, whereas *The Visible and the Invisible* represents its natural ontological extension.

Merleau-Ponty's epistemology considerably sets a limit of some conceptual tools employed in Husserl's phenomenology, such as those expressed by the notions of *intentionality*, *constitution*, *reflection*, *transcendental,* and gives stability to others such as those represented by the notions of *passivity*, *genesis*, *motivation*, *sedimentation*, noticeably extending their meaning. In many respects, concepts with a critical role in Husserl's phenomenological epistemology find in Merleau-Ponty a deeply different orientation. As Husserl's phenomenology, Merleau-Ponty's epistemological project is radically anti-reductionist and deeply anti-naturalistic.

> Scientific points of view, according to which my existence is a moment of the world's, are always both naïve and at the same time dishonest, because they take for granted, without explicitly mentioning it, the other point of view, namely that of consciousness, through which from the outset a world forms itself around me and begins to exist for me. To return to the things themselves is to return to that world which precedes knowledge, of which knowledge always speaks, and in relation to which every scientific schematization is an abstract and derivative sign-language, as is geography in relation to the countryside in which we have learned beforehand what a forest, a prairie or a river is. (p. IX)

However, differently from the Husserlian phenomenology, Merleau-Ponty's anti-reductionist attitude and anti-naturalism don't involve the suspension, or

* University of Florence

the bracketing, of the natural stance. In a different way, the anti-naturalism professed by Merleau-Ponty has the aim to recover and preserve the natural stance, as well as a space for the *pre-categorical* thought, within which the consciousness, by its nature and genesis, inhabits.

In other words, according to Merleau-Ponty, differently from Husserl, the naturalization and the natural stance don't follow the same way. The naturalization implies a process of conversion, that is, the translation of something derivative and secondary (for example the phenomenal and qualitative world) in something considered epistemologically basic and grounded (for example the world described by the physics). Instead, the natural stance reveals the necessity of an immersion in the broad context of nature, a process required if we want to give a full and authentic account of these "things" that phenomenology aims to describe from a morphological point of view.

The exclusion of the natural stance involves a description of the things very similar to that provided by a map, which is to a particular region what geography is to a landscape. Accordingly, the segregation of the natural dimension, in addition to the rebuttal of a natural attitude, risks to drain the content of the experienced thing, showing the image of a disembodied object, deprived of its flesh, that is a mere functional element with no depth.

In philosophy of mind, the rebuttal of the naturalistic stance, as well as the assumption of a natural attitude involve a departure from the supposition that the physical states, e.g., the neuronal states, are primary and irreducible elements. At the same time this involves a departure from a kind of *anti-reductionism* which, on the contrary, considers the states of consciousness as primary and irreducible, that is, as free elements independent from any natural position.

It is interesting to observe that the anti-reductionism, as stated by Husserl, implies the assumption of a reductive stance. Definitely, in certain respects, the concept of phenomenological reduction has a meaning contrasting the concept of reduction used in philosophy of mind. The phenomenological reduction requires giving up, or at least taking distance from, the natural stance (the scientific and object-oriented attitude) emphasized by reductionism in philosophy of mind.

However, as paradoxical as it may sound, the phenomenological reduction and the reduction in philosophy of mind share a critical aspect that justify, at least in part, their homonymy: both of them affirm the necessity of a radical

departure from the *natural stance* (in the case of phenomenology) and from the *manifest image* (in the case of philosophy of mind). Starting from this shared necessity, the phenomenological approach and the reductionism in philosophy of mind turn into two antithetical paths: the former establishes the priority of conscious experience and considers the physical states – the neuronal states included – secondary and derivative; while the latter establishes the priority of the physical states and considers the states of consciousness as derivative and according to some of its defenders not existing and illusory, therefore eliminable.

Assuming this point of view, the absence in Merleau-Ponty's works of a process of reduction – also of the phenomenological one – is perfectly clear. To endorse a philosophical project characterized by a radical anti-naturalism is not to deny the natural character of the consciousness. In this basic methodological distinction a critical change of paradigm can be summed up noticing that on the one hand the exigency of Husserl's phenomenology was that of disentangling the subject from the *world*, and that on the other Merleau-Ponty's phenomenology is concerned to completely immerge the subject in the world, restoring the natural *bilateralism* between thought and the environment that an original phenomenological description should always preserve.

The reflective subject of the Husserlian phenomenology, that is, the subject conceived as the condition of possibility, rather than the bearer, of an actual experience is the result of an analytic reconstruction and not of an original phenomenological description. Differently from this paradigm, in Merleau-Ponty's phenomenology there is no absolute priority for an impenetrable and objective reality, as well as there is no absolute priority for the idea of a subject conceived as a constitutive power, that is, as an invulnerable inwardness that can be reached through a backward walk.

Merleau-Ponty transforms the *correlative* analysis, typical of the Husserlian phenomenology within which the structure of consciousness is the basic element, in a *bilateral* analysis according to which both the subjective and the objective poles require a foundational priority. Accordingly, he extends the methodological approach from a perspective that privileges the external *frame* of the experience, to a perspective that fills that frame with an *actual content*.

In this view, the constitutive structure, or the reflective component, is progressively placed side by side with the domain of the unreflecting; the transparency of representation with the opacity of the feeling; the expressible character of the structured datum shows the relevance of the dumb, tacit,

unexpressed and inexpressible nature that the experience inexorably brings with itself.

This is a powerful change of perspective that makes it possible to transform puzzles in philosophy of mind (as in the case of the "question" of *qualia*), in "genuine" problems. On the other side, as noticed by Kuhn, the conversion of a puzzle in a problem becomes possible only when a change in the theoretical and conceptual background happens, a change that opens the door to a different definition of the problem and not to other solutions of the same puzzle.

This conceptual change is evident in the way Merleau-Ponty faces the problem of sensation as opposed to the puzzle of qualia. As it is well known, because of their subjective nature (intrinsic, private, and hardly reducible to a third person perspective) and their essentially qualitative character (direct, immediate, and so ineffable), qualia are considered in philosophy of mind the only and genuine *hard problem*. But Merleau-Ponty's phenomenology adds another trait, maybe the most important, to those that can be considered the standard features usually ascribed to *qualia*. Qualia are essentially and not accidentally associated to the subject's embodied dimension, that is, to the possession of a *lived body* contrasting the mere possession of a *physical body* (as in Descartes' philosophy). The introduction of the body establishes the role of the natural subject, that is, the role of the embodied, situated subject as regard to which both the notions of reduction in philosophy of mind and the phenomenological reduction appear to be inadequate.

On the other side, the introduction of the body determines an epistemological shift from the abovementioned *puzzle of qualia* to the *problem of sensation*.

> There are two ways of being mistaken about quality: one is to make it into an element of consciousness, when in fact it is an object for consciousness, to treat it as an incommunicable impression, whereas it always has a meaning; the other is to think that this meaning and this object, at the level of quality, are fully developed and determinate. (p. 6)

According to Merleau-Ponty, it is necessary to consider the question of sensitivity as a genuine *problem*: this is not a question concerning the possession of inert qualities or contents defined by well marked boundaries. Contrasting the identification of the notion of sensation with that of qualia assumed as a reply to external stimuli, the sensitivity is not something determined, instantaneous and detailed, but it is vague, ambiguous and

indeterminate. On the other side, for Merleau-Ponty, it is not correct to consider the domain of sensitivity as intrinsically formless and structureless except when a theoretical and meaningful system intervenes to check the rush and chaotic sphere of sensorial stimuli. This is the idea of a great part of post neo-empiricist epistemology, according to which, to be accessible the datum should be interpreted and embedded in a circle of hypotheses and background theories. On the contrary, according to Merleau-Ponty, the sensible datum is not tied to a theoretical and conceptual apparatus, but shows its own a proper structure, even if flowing and ambiguous.

The sensible field – that qualities inhabit – far from representing the immediate result of an external stimulus, or a mere reply to an external situation, depends on specific variables such as for example the biological sense of the situation. This makes the sensible experience a critical process analogous to that of procreation, or that of breathing and growth. Things are for Merleau-Ponty *flesh* and not mere *bodies*, they are not a mere extensions or bodily surfaces covered by specific qualities. Accordingly, sensations are not a mere reception of qualities but represent a living inherence, they don't offer inert qualities but active and dynamic properties characterized by a proper value related to their functional role in preserving our life.

> The pure quale would be given to us only if the world were a spectacle and one's own body a mechanism with which some impartial mind made itself acquainted. Sense experience, on the other hand, invests the quality with vital value, grasping it first in its meaning for us, for that heavy mass which is our body, whence it comes about that it always involves a reference to the body. (Merleau-Ponty, p. 60)

The identification between *qualia* and *sensitivity* derives from a process of alienation suffered by the concept of body that inevitably leads to the leveling off of both the notion of consciousness and the notion of experiential thing. Contrasting this view, the *embodied* thought becomes the result of a circular conception of experience and knowledge. This is a conception within which the experience assumes an *insight* that nor the Husserlian notion of plena, nor the notion of qualia in philosophy of mind, are able to show. In the first case because the former notion is too close to an extensional idea of the qualitative element. In the second, because the latter notion is too close to an empirical notion of sensible datum and to a physiologic and mechanistic interpretation of sensation.

The idea of sensation conceived as a filling quality and the idea of sensation assumed as the phenomenal and qualitative reply to an external stimulus, contribute to leveling out the domain of experience, draining and atrophying its own sense, that is, the idea of sensitivity as a living rhythm. A sensitivity that, in order to be understood, cannot be divorced from the analysis of the notions of body and embodiment, together with the awareness of the radical change of paradigm introduced by them.

Commentary

Neurophilosophy of Free Will

Henrik Walter

MIT Press, Cambridge (MA), 2001

Lorenzo Del Savio *

lorenzo.delsavio@ifom-ieo-campus.it

Neurophilosophy of Free Will frames the analytic debate about free will within current neurophysiological theories. The introductory chapter overviews several decades of discussion by listing three intuitions that should be accounted for by any eligible theory of free choice: freedom (ability to do otherwise), intelligibility (acting for reasons) and agency (being the source of our own choices). A moderate neurophilosophical *manifesto* is then outlined in the second chapter: knowledge of the brain should inform philosophy of mind. Chapter three eventually tries to meet these analytical and methodological *desiderata*: freedom, intelligibility and agency are extensively (though tentatively) naturalized by means of neuroscientific insights. Walter draws the conclusion that libertarianism should be rejected and free will explained by *natural autonomy*, a concept that should save *phenomena* and intuitions alike.

The relevance of Henrik Walter's book goes well beyond the issue that it explicitly addresses. His naturalization effort covers a wide range of traditional topics, from intentional content to the concept of a person. Many scientific theories that had been picked up were admittedly fairly hypothetical (p. 259) and so they still are. Thus the fate of Walter's specific proposals is open to scientific scrutiny. His main methodological point is nonetheless irreversible. Nobody would deny that philosophizing should be *conscious* of scientific developments. Walter claims it should be also *involved* in empirical inquiries: he urges for a *«bridge discipline between subjective experience, philosophical theorizing and empirical research»* (p. 125).

This review will focus on Walter's way of fulfilling his naturalistic program and, therefore, the first two chapters are let aside. It is nonetheless worth to

* University of Milan; IEO European Institute of Oncology

remember that, when the book was published, neither an extensive overview of the free will debate (chapter one) nor a plea for non-reductive physicalism (minimal neurophilosophy – chapter two) was yet commonsense. After more than a decade, Walter's three core proposals of naturalization instead deserve our attention. Part (1) presents Walter's ideas about chaos theory and free choices, part (2) deals with his naturalistic conception of brain content and part (3) outlines the link envisaged by Walter between the concept of a person and some neuroscientific insights about emotions.

(1) Walter's thoughts about freedom are organized in two sections: a *pars destruens* in which he argues that quantum physics is not relevant for the free will debate and a *pars construens* that borrows from chaos theory in order to dissolve the puzzle of freedom. The latter goes as follows: free choices require that, at some instant, more than one future is possible. Choices are bifurcating paths. Now, either the world is deterministic or it is not. If the former is the case, then there are no alternative paths by definition (van Inwagen, 1986) and it only seems there are. On the other hand, if indeterminism is true, then there are alternative paths at some point, but the choice is indeterminate. Therefore, which path is taken does not depend on anybody's decision (agency intuition) nor can it be explained by reasons (intelligibility intuition). The three core intuitions cannot hold simultaneously.

Penrose famously proposed to link agency and indeterministic phenomena of quantum mechanics to discard the second horn of the dilemma. According to Walter, linking agency and quantum phenomena has an obvious hurdle: agency is *prima facie* an organism level phenomenon, thus macroscopic, and macro-systems are practically deterministic. In fact, mainstream neurophysiologists take atomic and subatomic processes for granted and typically work on macromolecules (actually, dynamics of a huge number of them). To bridge the gap with the atomic level we would need an *amplifier* theory like Penrose's. A common criticism to this proposal points out that it rests on promissory notes about future physical theories. Walter's objection is rather that it is not even compatible with what we know about the brain: he gives compelling reasons to the effect that the brain is quite unaffected by atomic phenomena (p. 161).[1] Instead, we should focus on brain level phenomena rather than cell-level interactions.

[1] Those reasons are also independently interesting for who is concerned with inter-level reduction and levels of mechanistic explanations (see Darden 2008).

Compatibilist strategies often rest on a shift of meaning. Possibility of doing otherwise in the same circumstances is weakened and becomes possibility of doing otherwise *if one had wanted to*. Same circumstances are not identical, they are fairly similar – a move that resembles Lewis' thesis on identity across possible worlds. Would these counterpart-circumstances rescue freedom? Walter claims they would and try to explain why these circumstances are neurophilosophically relevant. In a nutshell, he suggests that brain network dynamics is likely to be chaotic and hence extremely sensitive to small fluctuations of parameters. «Using chaotic behavior, a cognitive system retains the option of reacting quickly, flexibly, and sensitively to relevant stimuli, changes in the environment, or ideas» (p. 182). These outcomes cannot be predicted despite their being wholly deterministic: epistemic indeterminacy suffices to account for the intuition that we could have done otherwise.

Yet an objection easily comes up: dependence on chaotic outcomes would end up in auto-epistemic indetermination. We would be astonished by our own decisions all the time. Not so, according to Walter: he takes a revisionist stance on decisions to blur the objection. «Decisions are not processes that occur at a point in time, they are events extended through time» (p. 183), indeed we would speak of mere reflexes – not choices – beneath of a certain time threshold (p. 184). Throughout the decision process, our cognitive system follows unpredictable trajectories, eventually resting down to a stable state: this is the decision. Nonetheless a major trouble remains: who controls the values that determine the trajectories? Walter admits to be again in a thicket, but a very different one indeed: the problem has shifted from availability of alternatives to agency. Walter's solution falls or stands with his naturalization of agency, a topic that is tackled towards the end of the book.

(2) Although the success of Walter's conception of free will depends mainly on later paragraphs, his treatment of intelligibility has several far-reaching philosophical consequences. The issue is spelled out in terms of acting for reasons and the latter is linked with the debate about intentionality. Millikan's ideas are then borrowed in order to explain intentionality of mental states in a physical world (neurosemantics). Intentionality is naturalized by adaptation and adaptations are explained by natural selection. Walter's step forward deals with the last concept: the scope of Millikan's teleosemantics gets wider to become *neurosemantics*. The explanatory power of selection is

stretched beyond the usual evolutionary time spans: selective-like processes in ontogenetic or even instantaneous times produce meaning *in the brain*.

Walter proposes that a physical state comes to be about a chunk of reality having a relational proper function. In Millikan's words, it is an intentional state that has been selected for conveying certain contents. Walter's specific contribution takes natural selection as a rather abstract schema and suggests that it applies to evolutionary times, ontogenetic times and even ultra-fast instants *phenomena*. The adaptive immune system is often the main example of a selective-like mechanism that produces specificity (antigens *recognition* – an intentional metaphor indeed) in ontogenetic times. The same token could well be true for the brain:

> among the constraints that support stability [of a brain structure] are not only complementary effects within the brain, but also interaction with the external world. [...] A temporarily stable state can be interpreted semantically because the stabilizing process is an adaptation. (p. 228)

An early proponent of this theory was not by chance Edelman (i.e., Edelman 1992), a Nobel-awarded immunologist.

A further dimension taken into account is subsequently the inter-subjective language, by mean of which content plays a physical causal role in the world. Causal networks including contents are *bona fide* physical interactions, nonetheless they might be *paraphrased* (p. 240) by reason talk in virtue of the selective history of their components. These intentional states, in the wording of Walter, supervene on physical structures and environmental surroundings.

Something of our intuitions about intelligibility has faded away (p. 243): an intentional state does not have causal power *as* intentional state but only as physical state, nonetheless it can be given an intentional content because of its history. It is noteworthy that the puzzle of free will vanishes even when one accepts this conclusion: here, reasons determine course of action only in a loose sense and the underlying physical process might well be indeterministic.

(3) Free choices belong to the physical (chaotic) causal network (freedom naturalized) and can be interpreted as reasons in virtue of the proper function of some physical state of the brain (intelligibility naturalized). Yet only a small subset of these reasons are recognized by a person as her own. «A compatibilist theory of agency must postulate that the determinants converging in a person are action of that person. In other words, it must be a theory about what makes

an executing instance a "self" or "person"» (p. 263). How can we naturalistically make sense of attributions of reasons to persons?

Frankfurt (1971) argued that identification with second order volitions is crucial: a person's will is free only if he is free to have the will he wants. Mention of freedom in the *definiens* would end up in a regress, but wholehearted identification with a volition guarantees that a second order will is authentically expression of a person. Yet we are not given further clarifications: Walter suggests neuroscience can provide some fruitful insights: indeed he claims that emotions play a pivotal role in this identification.

According to Damasio's work, while pondering, we simulate a counterfactual scenario by means of an imagined outcome of a choice and a correspondent body state representation. Crucially, body state representations get stabilized throughout the life of an individual, thus implicitly containing the past history of a person (p. 284). Frankfurt's regress of always higher-level volitions is stopped by emotional identification with a self-representation, Walter's naturalistic rephrasing of wholeheartedness. Only those volitions that are embedded in this emotional way are authentic. «Self-determined behavior is not a result of rational considerations, instead we learn to make clever and socially responsible decisions with the aid of our emotions» (p. 290).

Walter concludes summarizing his theory of natural autonomy:

> under very similar circumstances we could also do other than we actually do (because of the chaotic nature of our brain). This choice is understandable (intelligible – it is determined by past events, by immediate adaptation processes in the brain, and partially by our linguistically formed environment), and it is authentic (when through reflections loops with emotional adjustment we can identify with that action). (p. 299)

Whether this natural autonomy is compatibilist or hard-determinist won't concern us here: it all depends on our attitudes toward libertarianism (Kane 2001).

Still Walter's conclusions are open to conceptual as well as empirical challenges. Conceptually, the natural philosophy side of the debate loses its strength in Walter's treatment because there is no longer the issue determinism *versus* indeterminism in the foreground: natural autonomy is compatible with both metaphysics. This deflationary result instead brings authorship and the notion of agency at the core of the philosophical concern about freedom. Two main topics are therefore worth pointing out: consciousness and moral responsibility.

Although it is mentioned throughout the book, there is no extensive treatment of consciousness. Later works on free will turned to the topic claiming that consciousness lies at the very center of our conception of agency (and *hence* responsibility), namely Wegner's idea of consciousness as *emotion of authorship* (2002). Despite its frankly hard-deterministic framework, Wegner's proposal begins where natural autonomy ends. It shows how emotions, self and consciousness are entrenched.

A second topic is moral responsibility. Walter declares not to deal with moral theorizing because he wants to single out the pure metaphysical nucleus of the debate. I have argued that Walter's results turned out to be deflationary exactly as far as natural philosophy is concerned. Yet it is arguably not possible to clarify the concept of an agent without any link to responsibility, if not to explain why we have such a concept in the first place.

Aristotle's sea battle argument rested on logical worries. Medieval work on free will was carried out against a theological background. In modern times, the debate has shifted into a mechanistic framework and, perhaps more surprisingly, it has resurfaced even when statistical social laws have been discovered (Hacking 1990). Cognitive sciences have also been used as the last scenarios of this ancient battle (e.g., Libet 1985). Walter conceives the role of neurosciences more broadly. They do not simply challenge the pre-theoretical concept of free will. Rather, neurosciences might cast light on the notion of a person who is the author of her own decisions.

REFERENCES

Churchland, P. (2007). Neurophilosophy: The early years and new directions. *Functional Neurology, 22*(4), 185-195.

Darden, L. (2008). Thinking again about biological mechanisms. *Philosophy of Science, 75*, 958-969.

Edelman, G. (1992). *Bright Air, Brilliant Fire.* New York: Basic Books.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 11-25.

Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.

Kane, R. (2001). *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*, 529-566.

Van Inwagen, P. (1986). *An Essay on Free Will*. Oxford: Clarendon Press.

Walter, H. (2001). *Neurophilosophy of Free Will: From Libertarian Illusion to a Concept of Natural Autonomy*. Cambridge, MA: MIT Press.

Wegner, D (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Commentary

The Illusion of Conscious Will

Daniel M. Wegner
MIT Press, Cambridge (MA), 2002

Roberto Di Letizia *
r.diletizia@tin.it

In this Commentary, I intend first to introduce the philosophical discussion about the *conscious experience* of will advanced in Daniel Wegner's *The Illusion of Conscious Will*. Second, I will criticize his theory of conscious will as it does not solve the hard problem of consciousness. Furthermore, I will show how one of its keys concepts (namely, the distinction between nonvoluntary and voluntary actions) is a special case of mereological fallacy. In the end, I will refer to Dynamical System Theory (TSD) to suggest "to put into brackets" our natural attitude towards agency (inner mental states as causes of our actions), thereby introducing a more neutral framework to talk about natural agency as an emergent self-organizing behavior of nonlinear coupled systems.

1. The Argument of *The Illusion of Conscious Will*

According to Wegner, conscious thoughts are not the actual causes of our actions as they play no causal role in action-making processes. Instead, human behavior is caused by unconscious mental states at a subpersonal level, that is, the real causal sequence underlying human behavior involves massively complicated sets of mental mechanisms whereof the agent is not aware. Hence, if Wegner is right, free will is ruled out from our action-making processes: it is just an *illusion*.[1]

* University of Salento

[1] For Wegner, conscious will is an illusion as much as a magic trick (Wegner 2002, p. 27). The audience believes in a magic trick because the perceived causal sequence (i.e., «the set of events that appears to have happened») is the best and easy way to explain what is happened, when the real causal sequence, i.e., («the set of events the magician has orchestrated behind the scenes») is often more

Following Wegner, there are two different kinds of mental causation: (i) *apparent* mental causation, that is, *phenomenal will*, or "the feeling of doing something": conscious experience of will that is self-reported by an agent at a personal level; (ii) *real* mental causation (i.e., *empirical will*: the actual unconscious linkage between mind and action, namely an intricate set of physical and psychological processes at a subpersonal level).

From the distinction between these two kinds of mental causation, Wegner moves to explain why people believe that conscious mental states are the actual cause of their actions. He argues that «*people experience conscious will when they interpret their own thought as the cause of their action*» (Wegner 2002, p. 64).[2] Experience of will arises when the agent *infers* an apparent causal path between conscious thought and action whereas the actual causal path is not present in agent's consciousness: conscious thought and action are both caused by unconscious events. Thereby, experience of will is actually an *inference* about the cause of our actions which may be mistaken.[3] Most important, Wegner adds that this inference will produce the experience of conscious will only when the perception of the action satisfies three conditions: *priority*, *consistency*, and *exclusivity* of the thought about the action.

The priority principle claims that the experience of conscious will depends on «the timely occurrence of thought prior to action» since «causal events precede their effects, usually in a timely manner». Therefore, «thought that occurs too far in advance of an action is not likely to be seen as the cause of it» (Wegner 2002, pp. 70-71). The consistency principle claims that "the thoughts that serve as potential causes of actions typically have meaningful associations with the actions", which means that conscious thoughts occurring prior to the act must be semantically related to the latter. Accordingly, when people «think of one thing and do another – and this inconsistency is observable to them – their actions does not feel as willful» (Wegner 2002, p. 79). Finally, the exclusivity principle claims that the experience of will arises

complicated than the perceived sequence: «The illusion of conscious will occurs by much the same technique» (Wegner 2004, p. 653).

[2] This is Wegner's *theory of apparent mental causation* (Wegner and Wheatley 1999).

[3] As Wegner pointed out, it does not matter «how we are convinced that our thoughts cause our actions, it is still true that both our thought and action could be caused by something else that remains unobserved» (Wegner 2004, p. 655). Indeed, an experimenter can make arise the experience of conscious will so as to the subjects believe that they are controlling a perceived action though they are doing nothing actually (see Wegner 2002, pp. 74-78).

when the conscious thoughts are perceived as the exclusive cause of the action. Thus, the exclusivity that when conscious thoughts «do not appear to be the exclusive cause of the action" then people "experience less conscious will» (Wegner 2002, p. 90).

Nonetheless, according to Wegner, even though the experience of free will is a mere epiphenomenon it may work as a *mind's compass*. The feeling of doing is an *indicator* telling us something about the state of our own agency. Experience of conscious will inform us about the *authorship* of perceived causal sequences, whether what it is happening is or is not in our control: «conscious will is an emotion that authenticates the action's owner as the self»; – «This helps us to tell the difference between things we're doing and all the other things that are happening in and around us» (Wegner 2002, p. 327).[4] According to this hypothesis, we do not experience conscious will when the consequences of our actions do not satisfy one of these three conditions of causal inference: experience of conscious will is undermined by the «absence of any of these conditions» (Wegner 2002, p. 70). Consequently, an action is *experienced* in-control when the three conditions are satisfied as well as it is *experienced* out-of-control when one of those conditions is absent.

Wegner's philosophical main argument is based upon the explicatory distinction between the personal and the subpersonal level (e.g., Dennett 1969 and Stich 1978).[5] Conscious will belongs to the personal level whereas

[4] Wegner underlines that conscious will is an emotion, "an informative feeling" (i.e., Damasio's somatic marker; see also Damasio 1994).

[5] On the one hand, the personal level explains agent's behavior in the terms of their conscious mental states (i.e., desires, beliefs, plans, intentions). Accordingly, the causes of the behavior are conscious mental states such as intentions and purposes. For they are *teleological* explications. At this level, the content of mental states is *conceptual*, that is, (a) the subject is able to access *consciously* to it, (b) it is *compositionally structured* (i.e., inferentially integrated with other mental content, namely holistic), and (c) it is *semantically evaluable* by means of truth-conditions, truth-makers, and so on. Furthermore, the mental states are attributed to the whole *person* (the subject who perceives, believes, desires, acts). On the other hand, the subpersonal level explains agent's behavior in the terms of unconscious mental events (i.e., computational, functional, neurophysical states) attributed to domain-specific and informationally encapsulated cognitive subsystems, or modules. In other words, subpersonal states are attributed only to an *anonymous* part of a person: the brain. Finally, the content of subpersonal states is *nonconceptual*, which means that (a) the subject *cannot* access consciously to it, (b) it is *unstructured* (i.e., inferentially isolated), and (c) it is *non-semantic* (i.e., it does not have truth conditions). Finally, sensory inputs, neural events and motor outputs are connected by *causal factors*. For subpersonal explanations are not teleological, but *mechanistic*. (For a debate on the relationship between these two levels, see Clark 2003 and Bermúdez 2003; for a critical point of view

the actual causes of our actions lie on the subpersonal one. I suggest that the problem of free will (namely, how can something like the free will exist in a causally determined universe?) emerges when human behavior is explained at the subpersonal level. At this level, mental phenomena are explained as *mechanisms*[6] whose function is to connect some sensory inputs to some motor outputs. Most important, a mechanism is causally determined, its operations are always initiated or maintained by an external cause and the state of each component depends on the operations of another component. Now if we think, as Wegner does, that action-making process is sustained by psychological mechanisms, then free will cannot play any causal role in them. For the free will is *not* a mechanism but it is an uncaused cause which cannot be caused by any external cause.

2. Wegner's Anti-Libertarian Incompatibilism

The current philosophical debate on free will shows two opposite views: compatibilism and incompatibilism (Watson 1982). On the one hand, according to compatibilism, causal determinism does not rule out the free will. On the other hand, according to incompatibilism, free will is not consistent with the causal determinism. Furthermore, there are two different kinds of incompatibilism: libertarianism and anti-libertarianism. Libertarianism claims that free will exists and, consequently, causal determinism must be false. On the contrary, anti-libertarianism claims that free will does not exist *because* causal determinism is true. With the respect of these sketched framework, Wegner's account of free will belongs to the anti-libertarian incompatibilist view.

First, it endorses incompatibilism because free will and causal determinism are «incommensurable» (Wegner 2002, p. 322). Free will is conceivable only as an uncaused cause, which should be *«unresponsive to any past influence»* and derives from agent's ability «to do things that do not follow from anything» (Wegner 2002, p. 323). However, if causal determinism is true, then

on personal/subpersonal distinction, see Hurley 1998, pp. 29-54; Bennett and Hacker 2003, pp. 68-107).

[6] A mechanism is a structure performing a function in virtue of its components parts, component operations, and their organization (e.g., Bechtel and Richardson 1993).

everything is caused by something else, and the concept of uncaused cause is not acceptable

Second, Wegner's account is anti-libertarian as he thinks that free will cannot be integrated in a rational theory of human action. Indeed, free will as an uncaused cause is caused by nothing, neither by the agent. For this reason, free will can act only randomly. It follows that none is able to control their own actions, and that free will deprives the agent of any causal power on his/her own actions. Instead, Wegner claims that only causally determined psychological mechanisms can provide us to an effective theory of human action: «free will is not an effective theory of psychology and has fallen out of use for the reason that it is *not the same kind of thing* as a psychological mechanism» (Wegner 2002, p. 324).

As anti-libertarian incompatibilist approach, Wegner thus proposes eliminative view about free will concept. Indeed, if free will cannot describe the actual psychological mechanisms causing human action, then it can be ruled out from the psychological vocabulary.[7] Wegner advises a paradigm shift in the analysis of free will from intentional psychology to cognitive neuroscience. Depending on this ungrounded concept, the debate between determinists ("robogeeks") and free-willers ("bad scientists") is futile and ill-posed as it depends on the concept of free will, but if we eliminate this concept, we eliminate the debate as well. Instead, we should not look for a neural surrogate of free will because free will is conceptually wrong. Rather, we have to study two distinct phenomena: mechanisms of action-making and feeling of doing. The former consists in causally determined unconscious thoughts that are the actual causes of our actions. The latter (namely, conscious will) is just a kind of feeling, a perception detecting whether an action is in control or out of control:

> Whether we embrace the illusion of control or reject it, the presence and absence of the illusion remain useful as clues to what is real. Just as the experience of will allows us to know what we can control, the lack of this feeling

[7] We can outline three reasons to eliminate free will following Paul Churchland's eliminative materialism (Churchland 1981, pp. 75-76): (i) free will suffers explanatory failures on epic scale, it explains only some aspects of human actions but it is not able to solve many others issues (e.g., How can an uncaused force exist in a deterministic world?); (ii) free will has been stagnant for a long time as compatibilist and incompatibilist views still show the same unsolved problems (e.g., the problem of self-control); (iii) free will explanations are not reducible to neuroscience because they involve uncaused processes whereas brain's processes are mechanistic.

> alerts us to know what we *can't* control, what surely exists beyond our own minds. (Wegner 2002, p. 333)

3. The Hard Problem and the Mereological Fallacy

I suggest Wegner's account on conscious will does not succeed to solve two problems, namely the *hard problem* and the *mereological fallacy*.

First, the hard problem is the problem about the conscious experience, that is, *why* something like conscious experience exists (Chalmers 1996, ch. 3). For cognitive sciences, consciousness is a hard problem because whilst psychological states can be reduced to functional or computational states, the consciousness resists to any reductionist attempt. Indeed, two subjects may be functionally identical even though only of them has a conscious experience. Therefore, psychological explanations are *blind* about conscious experience insofar as they do not distinguish a subject who has a conscious experience from a subject who has not (e.g., a zombie, a robot). Now, Wegner's account exposes conscious experience as a detector of authorship (a "mind's encompass") about our actions. Nevertheless, robotics shows us that some embodied agents are able to control and to detect whether an action is self-performed or not without conscious experience. This is recognized by Wegner himself when he writes that even a robot may have conscious will if it was able «to keep track of what it was doing, to distinguish its own behavior from events caused by other things» (Wegner 2002, p. 340). However, if conscious experience is not necessary for an authorship detector, then we are not explaining *why* in human beings the former supervenes on the latter. In other words, the hard problem is still there: *why* does the conscious experience of will exist if an embodied agent is able to detect the authorship about its own actions without conscious experience?

Second, Wegner's account rules out the distinction between in-control and out-of-control actions. Indeed, according to Wegner, conscious experience of will is a kind of knowledge, it is nothing else than an inference about the causes of our actions. As a result, the *voluntariness* experience of our actions is an illusion, for the subject does not actually control his/her own actions. However, if the subject cannot control his/her actions at all, *nonvoluntariness* experience is an illusion as well. For non-voluntariness is not a matter of fact, rather it is an *epistemic* instance that informs the subject when his/her knowledge about the cause of our actions is wrong. Accordingly, the

distinction between voluntariness and non-voluntariness is not an ontological instance but it depends on subject's epistemic structure: actually there are not real things like voluntariness and nonvoluntariness actions.

I suggest that the concepts of being-in-control and being-out-of-control are related to folk psychology inasmuch as they imply a substantial Self (i.e., a central controller) enabled of controlling its behavior by means of conscious thoughts, but this is exactly what Wegner denies. As *personal* categories, at a subpersonal level voluntariness and nonvoluntariness have not place, for there is not a *person* enabled of controlling his or her behavior. The point is that at the subpersonal level, we have only loop circuits or recurrent networks wherein the events are transformations of state vectors, whilst we can see voluntary or non-voluntary actions only if we *interpret* these subpersonal events as result of a conscious Self which is endowed with contentful mental states. Consequently, Wegner makes the "*mereological fallacy*" (Bennett and Hacker 2003, p. 73). He applies psychological predicates, which are attributable only to human beings as whole (i.e., a Self) to subpersonal processes and states.

I suggest that Wegner's theory on conscious will make the mereological fallacy as they contain descriptions which are encapsulated in the human observer's "cognitive domain" (e.g., Maturana and Varela 1980). The cognitive domain is nothing else than the observer-centred theory which describes the cognitive system's behavior in terms of inner mental states (i.e., propositional attitudes, informational states, inner representations). Thereby, the challenge is to explain why agent's behavior shows recurring patterns of activity, which constitute his *personality*, without any reference to observer-centred descriptions. In order to provide a naturalist account of agency, we ought "to put into brackets" our natural attitude towards the agency, which posits (un)conscious mental states as the causes of the behavior, and to address to a more neutral framework (namely, non-observer-centred).

4. Agency in Motion

Dynamical Systems Theory (TSD) may be a powerful framework to explain natural agency.[8] Self-organizing complexity is a powerful tool for

[8] Dynamical system's state evolves in real time and may show significant nonlinearities (i.e., it is often discontinuous, or disproportional, and hardly predictable as well). Dynamical system's state (i.e., *instantaneous physiologic state*) changes continuously in time plotting *trajectories* in phase

understanding psychological systems (e.g., Piers *et al.* 2007) as well as agency without personal concepts such as voluntariness and nonvoluntariness. The brain is a self-organizing nonlinear system coupled with the environment. Thus, the behavior of the system depends on many variables concerning both the nervous system and the environment.[9] As any nonlinear dynamical systems, psychological systems will show a self-organizing dynamics (i.e., phase transitions, attractors). In this sense, *personality*, which depends on recurring patterns of the agent's activity, is the spontaneous dynamics of the brain-environment system:

> There is no unitary 'ego' or 'self' that directs what we do. Instead, the spontaneous activity of neurons and groups of neurons, in continual transaction with the environment, is associated with the complex emergent activity we call *personality*. A complete description of personality therefore should involve neuroanatomy, neurodynamics, environment, and functioning.

space. Most important, their dynamics may show *phase transition*, *attractors* (i.e., regular patterns of activity which may be periodic, quasiperiodic or chaotic), and *repellors* (i.e., unstable configurations of a system which tends to "avoid" them). Nonlinear dynamical systems encompass *chaotic*, *complex*, and *self-organizing* systems. A chaotic system has two proprieties: i) it is *sensitive to initial conditions*; ii) its behavior is *unpredictable* over a long time level though it is strictly deterministic. A complex system is composed by a network of heterogeneous parts that interact nonlinearly in order to produce an emergent global behavior. A self-organizing dynamical system has no internal or external program that directs its functioning, though its behavior can produce recurrent patterns of activity. Biological systems, such as a brain, are complex, dynamic, nonlinear, chaotic and self-organizing systems (e.g., Kelso 1995). (For a general introduction to TSD see, also Stewart 1990).

[9] Thus, this dynamical account is clearly externalist. Indeed, when we talk about the object of study of cognitive science, we can be internalist or externalist. Rougly, internalism claims all cognitive processes and states are encapsulated in the head of the subject, so that it proposes a methodological solpsism: the behavior of the subjects can be explained referring only to the internal processes and states occurring in their own brains. Externalism, instead, claims cognitive processes and states extend and encompass features of the physical and social environment (e.g., Clark and Chalmers 1998; Wilson and Clark 2009). How is it possible? Part of answer lies in the premise of TSD: brain and environment are nonlinear coupled systems (e.g., Van Gelder 1998). Indeed, as Tony Chemero and Michael Silberstein have pointed out: «Dynamical systems theory is especially appropriate for explaining cognition as interaction with the environment because single dynamical systems can have parameters on each side of the skin. That is, we might explain the behavior of the agent in its environment over time as coupled dynamical systems, using [...] coupled, nonlinear equations» (Chemero and Silberstein 2008, p. 14). In other words, the state changes of the brain depend on changes in the external environment as much as the changes in the external environment depend on the changes of the brain. For it is important for cognitive modeling to track causal processes that cross the boundary of the individual organism as it is to track those that lie within that boundary.

(Grigsby and Osuch 2007, p. 42)[10]

From this dynamical standpoint, personality depends on the dynamics of the brain-environment system which may exhibit basins of attraction (namely, a region of states wherein more attractors are placed) and repellors. As the variables of the systems are distributed between brain and environment, a small change in the brain activity, or in the environment, may provoke a global evolution in the whole system changing its basin of attraction. Some basins of attraction are "stronger" and more stable than others insofar as they can be changed only by altering order parameters deeply. In fact, unlike "weak" attractors and repellors, the change of a "strong" attractor requires much energy and time.

Agency is a self-organizing capacity of the system of altering its own state by engaging in certain actions. In fact, nonlinear dynamical systems are well-know for the *circular causality* (Kelso 1995, pp. 8-9), that is, their own states are able to alter the order parameters in order to alter their own states. The significant propriety of self-organizing systems is the capacity of adapting their spontaneous dynamics according to the changes of order parameters. Consequently their dynamics is context-sensitive, for its evolution depends on the changes of order parameters.

Order parameters may be changed by some performed actions that provoke a phase transition switching the basin of attraction. Depending on gravity force of the basin of attraction, the agency is a continuous fuzzy process that may require time for changing dynamics:

> those acts that require greater alterations from habitual patterns of behaving require more agency (viz., greater deliberate effort) than those that represent repetitive behaviors with strong attractors and high probability of occurring. (Grigsby and Osuch 2007, p. 64)

[10] As nonlinear result of a complex dynamical system, person's behavior is determined by many factors. Some of them operates at the cellular level (e.g., membrane permeability and ion channel conductance, blood glucose level, concentration of neurotransmitters), others operate at a neurodynamical level (e.g., emotional state, motivational status, pain or discomfort, level of energy/fatigue, level of arousal), others reflect physiological state such as the sleep-wake cycle or neuroendocrine influences (e.g., cortisol, testosterone, progesterone, adrenaline), and still others are environmental features such as ambient temperature, level and type of sensory stimulation), or the presence or absence of certain people (e.g., parents, enemies) (Grigsby and Osuch 2007, p. 42).

5. The dynamics of Free Will

Surprisingly, the dynamical agency view may be consistent with free will, although self-organizing systems are deterministic.

Firstly, we can reshape the concept of *autonomy* or *libertas spontaneitatis*. According to the classic view, an action is not free-willed if it is heterodetermined. However, from the standpoint of TSD, there are not distinction between endogenous and external causes since brain and environment are a whole system. Environmental and cerebral factors are equals: there is not an inner Self separated to an outer environment. As Maturana and Varela have pointed out: organism and environment are *structurally coupled* (Maturana and Varela 1980).

Even though the distinction between autonomy and heteronomy does not make sense, we are nevertheless able to reshape the concepts of voluntariness and nonvoluntariness without a central controller. Control is not a dichotomist propriety but it is conceivable as continuous gradual process so as to a system can have more or less control. As a consequence, systems with more control are those that are able to change easily their basins of attraction, whereas systems with less control are those that are not able to change basins of attraction even though the order parameters have been altered by their own actions. Indeed, strong attractors are invariant respect to initial condition, for this reason the systems with strong attractors are not really responsive to environmental changes. If so, the behavior's stability is not a synonymous of control, but of out of control. Instead, the random, chaotic or unstable activity of the brain is warranty of control because this kind of activity allows the brain to be in «a state of maximum responsiveness» (Freeman 1995) so that it is «poised on the brink of instability where it can switch flexibly and quickly» (Kelso 1995, p. 26). Inasmuch as the brain has and shifts multiple co-existent attractors, which can be competitive or cooperative, the dynamics of the brain-environment system is "metastable". Accordingly, the agents experience loss of control when some attractors are stronger than others so as to they cannot change the behavioral patterns. This does not mean that the Self is weak but that the Self *is* the intrinsic dynamics of a dynamical system (namely, an autopoietic unity) which is able to self-produce and self-regulate its own processes. In this sense, the behavior is what the organism *does* when it engages the world by actively regulating its exchanges with it (e.g., Di Paolo 2005). As autopoietic system, the organism's behavior has the only purpose of maintaining its intrinsic

dynamics in a range of state's values. In other terms, natural agency is the *regulation* of the organism's intrinsic dynamics which is enacted by itself in order to maintain the state's variables in a certain range of state's value.

Second, a dynamical view of agency can partially preserve *libertas indifferentiatae*. How could agent's ability of "doing and choosing otherwise" be consistent with a deterministic view? First of all, *libertas indifferentiae* depends on a decision-making process. Decision-making process can be understood using theory of chaos as a trajectory of a system unfolding in real time: beginning at an unstable state, "visiting" various places in phase space and finally moving toward a stable state that corresponds to the nonchaotic, or chaotic, basin of attraction (Walter 2001, p. 185). Hence, decision-making process is a point of instability (namely a 'bifurcation') into a phase space where the behavior of the system is unsteady and fluctuating so as to it could take either of two directions until it settles down in a steady state. In this sense, the agent, as chaotic system, could have chosen or done otherwise. Furthermore, initial conditions does not causally determine the behavior of the system, rather a chaotic system is more or less sensitive to some changes and variables. Therefore, the switch from chaotic to stable behavior can be achieved by altering a single order parameter. Changes of order parameters can only increase or decrease the probability of occurrence of a behavior where attractors and repellors are only behaviors with a high or a low probability of occurrence. Hence, the ability of "doing and choosing otherwise" is a continuous gradual process as some changes of order parameters can increase or decrease the probability of occurrence of some behaviors reducing subject's agency.[11]

[11] For instance, consider a subject affected by a tumour of adrenal gland provoking an overproduction of hormones. This disease causes features of his personality such as aggressive mood. Suppose he has killed his wife when he had a fit of anger: could he have done or chosen otherwise? Probably he did, but an alternative behavior had a low probability of occurrence because his tumour has changed some order parameters (i.e., hormones level in the blood). So his agency has been reduced by order parameters' alterations that have increased the probability of occurrence of a aggressive behavior, seen as a strong attractor. Degree of agency depends on system's chaoticness: the initial conditions (i.e., order parameters) can increase, or decrease, the degree of stability of a system. Thus, when they increase the instability of the system, more the system will have the control of its behavior and the capacity of could have done otherwise (i.e., of switching from a steady state to another one).

6. Conclusion

The aim of this Commentary was to philosophize Wegner's theory of conscious will. I have suggested some philosophical implications and a proposal of solution. First of all, I have introduced the main philosophical argument assumed by Daniel Wegner in *The Illusion of Conscious Will*, namely the distinction between personal and subpersonal level of explanation. Wegner's theory of apparent mental causation claims the existence of conscious and unconscious mental states, where only the latter are the actual causes of our actions. Second, I have situated Wegner's account of free will in the current philosophical debate in which it may be seen as a form of anti-libertarian incompatibilism. Third, I have criticized Wegner's theory as it does not solve the hard problem of consciousness. Moreover, it fails to distinguish in-control and out-of-control actions. The reason of this misunderstanding is in the mereological fallacy: Wegner applies psychological predicates (voluntariness and nonvoluntariness), which are attributable only to human beings as whole (i.e., a Self, central controller, executive program), to subpersonal processes. Fourth, I have proposed as neutral framework the Dynamical System Theory (TSD) which may allow us to "put into brackets" our natural biases concerning agency. I have suggested that agency is the self-organizing capacity of a nonlinear dynamical system of altering its own state by engaging in certain actions without controller by adapting their spontaneous dynamics according to the changes of order parameters. Finally, I have sketched up two dynamical accounts of the concept of agency in order to reshape both the concepts of *autonomy* or *libertas spontaneitatis* and of *libertas indifferentiatae* by means of tools of TSD.

References

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. London: Blackwell.

Bermúdez, J. (2003). *Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States*. In Y. Gunther (Ed.), *Essays on Nonconceptual Content*, (pp. 183-216). Cambridge, MA: MIT Press. [1995]

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Chemero, T., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science, 75*(1), 1-27.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy, 78*(2), 67-90.

Clark, A. (2003). *Connectionism and Cognitive Flexibility*. In Y. Gunther (Ed.), *Essays on Nonconceptual Content*, (pp. 165-182). Cambridge, MA: MIT Press. [1994]

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10-23.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Avon.

Dennett, D. (1969). *Content and Consciousness*. London: Routledge.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429-452.

Freeman, W. (1995). *Societies of Brains: A Study in the Neuroscience of Love and Hate*. Hillsdale, NJ: Lawrence Erlbaum.

Grigsby, J., & Osuch, E. (2007). Neurodynamics, state, agency, and psychological functioning. In C. Piers, J. Muller & J. Brent (Eds.), *Self-Organizing Complexity in Psychological Systems*, (pp. 37-82). Lanham, MD: Rowman & Littlefield Publishers.

Hurley, S. L. (1998). *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordecht: Reidel Publishing Company.

McGinn, C. (1989). Can we solve the Mind-Body Problem? *Mind, 98*, 349-366.

Piers, C., Muller, J., & Brent, J. (2007). *Self-Organizing Complexity in Psychological Systems*. Lanham, MD: Rowman & Littlefield Publishers.

Stewart, I. (1990). *Does God Play Dice?* London: Penguin Books.

Stich, S. (1978). Belief and subdoxastic states. *Philosophy of Science, 45*, 499-518.

Van Gelder, T. (1998). The dynamical hypotesis in cognitive science. *Behavioral and Brain Sciences, 21*(5), 615-628.

Walter, H. (2001). *Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy*. Cambridge, MA: MIT Press.

Watson, D. (1982). *Free Will*. New York: Oxford University Press.

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wegner, D. (2004). Précis of *The Illusion of Conscious Will*. *Behavioral and Brain Sciences, 27*(5), 649-659.

Wegner, D., & Wheatley, T. (1999). Apparent mental causation: Sources of experience of will. *American Psychologist, 54*, 480-491.

Wilson, R. A, & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge Book of Situated Cognition*, (pp. 55-77). Cambridge: Cambridge University Press.

Commentary

The Volitional Brain

B. Libet, A. Freeman and K. Sutherland (Eds.)
Imprint Academic, Thorverton (UK), 2004

Elisabetta Sirgiovanni *
elisirgiovanni@hotmail.com

First published in 1999, as a collection of 18 influential papers from two important issues (No. 8-9) of the *Journal of Consciousness Studies*, and reprinted in 2004, *The Volitional Brain* results a work in what would have been called, at least since 2002, "neuroethics" (see Illes 2006). The neologism "neuroethics" does not appear in the book. As a matter of fact, even though first mentioned by A. Pontius on *Psychological report* in 1993, the term rapidly imposed itself only after a series of meetings in Europe and United States in 2002 producing a general agreement of a new burgeoning disciplinary field on brain research related to ethical and moral issues (just think about the publication of the proceedings *Neuroethics: Mapping the Field* by Dana Foundation in 2002). Given two general approaches to neuroethics (Roskies 2002), the *ethics of neuroscience* and the *neuroscience of ethics*, the book turns out to be a work in the neuroscience of ethics and thus privileges cognitive neuroscience, instead of philosophical bioethics, as a framework for ethical theory. This approach has been recently developed and has produced a number of international works among which this book can be considered a real classic.

In its four sections (Neuroscience, Psychology and Psychiatry, Physics, Philosophy), followed by Comments, the book discusses the relevance of neuroscience research for free will debate pertaining to the different theoretical areas. The various chapters arises as comments of the editor Benjamin Libet's results in 1985 showing automatic unconscious brain processes, preceding the awareness of a decision, as responsible of human volitional behavior. Although the reference to the unconscious with regard to

* University of Rome "La Sapienza"

free will could recall the debate about Sigmund Freud's psychic determinism, it should be clear for outsiders that Libet's intent was referring to non-conscious mental events – in other words, completely inaccessible to consciousness – differently from what stated by Freud through the object of psychoanalytic therapy and properly expressed by himself as "sub-conscious" (Ostrowick 2007). Thus, according to neuroscientific findings,the problem of free will becomes even more worrying. Are we free? Are we the authors of our volition? The way this book faces the subject is surely not a traditional one.

Traditionally, free will has been considered a problem of over-causation between human volitional causation and deterministic one (that of God foreknowledge or physical laws). The spectrum of traditional responses has framed the debate concerning the relationships between free will and determinism. In other words, whether free will and determinism were mutually exclusive opposite (incompatibilism) or not (compatibilism). Incompatibilism provides solutions among libertarianism (indeterminism), hard determinism (free will illusion) or skepticism (randomness). The book covers all positions well.

The aim of the book is to highlight Libet's findings thanks to more recent scanning techniques as PET and fMRI (see Chs. 1, 2, 3). Despite what Libet's results might appear at a first sight, the book maintains an equilibrium between the compatibilist alternatives (e.g., Gomes, Ch. 5; Clark, Ch. 18) and Libet's work (Ch. 4): it just lightens the idea of free will (as a consciousness veto over volitional activity) but does not jettison it. Contrary to all expectations for a neuroethical text, even anti-materialistic positions (see Chs. 17, 13, 8, 11) or suggestions from Eastern cultures and meditation traditions (especially Buddhism, see Chs. 8, 6, 7, 14) are presented. These two groups of articles, which respectively prefer non-physical mental forces as a solution even in clinical contexts (see Schwarz, Ch. 8, on OCD) or offer an "ambiguous phenomenology" (Libet *et al.*, Introduction, p. XIX), are a fault for a book that pretends to be neuro*scientific*. Moreover, the discussion on mind-body relation interestingly involves constraints coming from physics: laws of nature conception (Hodgson, Ch. 12), quantum theory (Stapp, Ch. 9), conservation law (Mohrhoff, Ch. 10), time (Lanier, Ch. 15). And finally it turns to law and compares free will to the problem of the power and penal responsibility (Chs. 16, 17).

The non-traditional way according to which the book presents the subject is referring to the reducibility of folk psychological notion of volition (and

choice) to brain processes. Admittedly free will has two components to be showed. Obviously free will is something dealing with freedom and will. So free will contains a metaphysical component (freedom) and a psychological one (will). As freedom has been traditionally contrasted by referring to physical laws (scientific determinism) and the mental event of volition to neurological causation (mind-brain problem), free will can be regarded as a question of reducibility of higher-level causal processes and explanations to lower-level ones. Accordingly, freedom and volition are two common sense intuitive notions related to the scientific conception of the world.

Nevertheless the traditional philosophical debate on free will has attributed a low value to the volitional component, so that the entry "free will" has been explained as «the conventional name of a topic that is best discussed without reference to the will» (Strawson 1998, p. 743). What I am going to discuss here is whether such a book, which deals with the *volitional brain* in order to propose what explicitly declared in the subtitle as a "neuroscience of free will", can genuinely represents a contribution to the free will debate. Or rather, whether (1) investigating volition is relevant to free will, and (2) neuroscientific findings can challenge or inform our notion of free will (see Roskies 2006).

First of all, there are three kinds of freedom: social freedom, which is conceived as a relation between an agent, an action and a power and sounds like "I'm free to do X with regard to P if P cannot oblige me to do it or prevent me from doing it"; freedom of action, which is a relation between an agent and an action in the sense that " I'm free to do X if I am able or I have a chance to do it"; freedom of will, which made the philosopher Jean-Paul Sartre (1943) seeing humans as "condemned to be free" and corresponds to something like "I could have acted in other ways, as I act on the basis of reasons, that is, I am the author of my decision". Only the third kind of freedom pertains to free will as the will is the entity that needs to be characterized as free. Questions at the end of the second paragraph can be hence reformulated as following: What is it to act (to choose) freely? What is it to be morally responsible for one's actions (or choices)?

It should be mentioned that a psychological conception of free will as self-determination is the basis of penal law theory. The core of imputability in Western penal codes is the volitional character of a criminal action, independently from how free will is intended as a metaphysical notion, namely

its reducibility or not to physical causation. Therefore, volition is at least an important component in the way we are ordinarily involved in the matter.

Nevertheless there is an argument according to which neuroscience is not in a position to undermine our intuitive notion of free will, and consequently that of moral (and then penal) responsibility. The argument focuses on the fact that problems on these notions exist independently of neuroscientific advances and depend on the existence of external forces such as God or nature (Roskies 2006). Neuroscientific inquiry is a matter of discovering mechanisms underlying cognitive phenomena (Bechtel 2008 and Craver 2007), while the problem of free will is a metaphysical problem that regards the deterministic (or indeterministic) nature of the universe. It is true that a naturalistic investigation of the wider problem concerns more physics than neuroscience. But intuitive concern on free will maintains that human agency requires freedom whereas mechanisms behave deterministically and that is why volitional brain mechanisms have been recently called into question. Regardless whether or not the universe is deterministic, however, neuroscience aims to show at best whether the brain is. So even if this work cannot give an answer to the wider metaphysical problem, it is still an important direction of inquiry.

Contrary to what people think, mechanism and determinism are not the same thing. A view of ourselves as biological mechanisms should not necessarily undermine our freedom. There are various ways to escape the problem. For example, recent neuroscientific accounts claim that «freedom is not freedom from causation, but the freedom of a system that is directing its own engagement within its environment» (Bechtel and Abrahamsen 2007, p. 63).

What Libet's results showed is that people are not actually *conscious* of their decisions. So these experiments focus on the relation between consciousness and free action within the brain. As a matter of fact we think of free will as *self*'s ability to choose whether or not to act. There are arguments against this view and against the link between awareness and decision (for a discussion, see Mele 2005). Apart from the metaphysical framework we choose, our intuitive notion of free will regards our *feeling* of control on our decisions and actions, not the control itself. And this accounts for neuroscientific inquiry. For example, literatures has presented contradictory experiments showing folk conception on free will both compatibilist and incompatibilist depending on the circumstances (Roskies and Nichols 2008).

Therefore even if cognitive neuroscience cannot give an answer to the question of freedom with regard to determinism, it can evidence other factors, which may inform our evaluations on freedom and responsibility. These factors are features of the functioning of mechanisms of choice and decision-making underlying folk psychological processes we refer to when we attribute freedom or responsibility to agents. Independently from the deterministic or stochastic nature of these mechanisms, their understanding corresponds to such essential attribution.

We usually count on our intuition of free will, we make use of it in our ordinary lives and in legal contexts. Recent titles testify that the interest in free will has come back again thanks to neuroscientific discussions introduced in books like this. Even though each paper should be judged separately from the others and some of them might result worthless if we refer to present debate, this text should be read as a precursor. It is a topical work facing the problem of a neuroscience of free will.

REFERENCES

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2007). Explaining human freedom and dignity mechanistically: From receptive to active mechanisms. *The Journal of Philosophical Research, 32*, 43-66.

De Caro, M., Lavazza A., Sartori G. (Eds.) (2010). *Siamo davvero liberi? Le neuroscienze e il mistero del libero arbitrio*. Torino: Codice Edizioni.

Illes, J. (Ed.) (2006). *Neuroethics: Defining the Issues in Theory, Practice and Policy*. Oxford: Oxford University Press.

Mele, A. (2005). Action theory meets neuroscience. Paper at the International Conference on Intentionality, Deliberation and Autonomy, University of Siena, 11st-13th March 2005: http://www.unisi.it/eventi/practical_philosophy/paper/Mele.pdf.

Morris, S. G. (2009). The impact of neuroscience on the free will debate. *Florida Philosophical Review, IX*(2), 56-77.

Ostrowick, J. M. (2007). The timing experiments of Libet and Grey Walter. *South African Journal of Philosophy, 26*(3), 9-26.

Roskies, A. L. (2002). Neuroethics for the New Millennium. *Neuron, 35*(1), 21-23.

Roskies, A. L. (2006). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences, 10*(9), 419-423.

Roskies, A. L., & Nichols, S. (2008). Bringing moral responsibility down to earth. *Journal of Philosophy, 105*(7), 371-388.

Strawson, G. (1998). Free will. In *Routledge Encyclopedia of Philosophy* (pp. 743-753). London: Routledge.

Commentary

Living Without Free Will

Derk Pereboom
Cambridge University Press, Cambridge, 2001

Giuliano Torrengo *
giuliano.torrengo@unito.it

The thesis that we utterly lack free will and thus we are not morally responsible for our actions looks difficult to reconcile with the basic features of our ordinary experience. In his influential book, Derk Pereboom argues in favor of it, as the most rational view on agency on the market. In this commentary, I will discuss his view and make few remarks on the connection between certain problems in the philosophy of agency and problems in the metaphysics of time and persistence.

Remember the general framework of the debate over free will and moral responsibility. If someone is morally responsible for her actions, she can deserve praise or blame for them. But moral responsibility requires freedom, since if someone is not free, she cannot deserve praise or blame. The *compatibilist* thinks that being free in the sense required for moral responsibility is compatible with physical determinism. We can be free even if our actions are determined by all what has happened in the past, given the actual laws of nature. The *incompatibilist* disagrees and thinks that if determinism is true, we can't be free in the sense that matters for moral responsibility, and thus we cannot be morally responsible. Traditionally, there are two varieties of incompatibilism. The first is *libertarianism*, who maintains incompatibilism and denies determinism. According to the libertarian, we are free agent in a indeterministic world. Therefore, we are morally responsible. It follows that praise and blame can be rationally justified reactions to human actions. The second one is *hard determinism*. According to the hard determinist, incompatibilism and determinism are both true. Thus, we are agents who lack free will and who inhabit a deterministic world. We lack moral

* University of Turin

responsibility and judgments of moral praise or blame are always irrational – they only seem to be justified.

Pereboom defends a form of incompatibilism – which he calls *hard incompatibilism* – that is neither libertarianism nor hard determinism, although it is clearly closer to the latter than to the former. This is true not only because hard determinism is subsumed under hard incompatibilism: being a hard determinist is a way of being a hard incompatibilist, although not always the other way around. But more profoundly because Pereboom's incompatibilism, in a sense, embodies the gist of hard determinist, without entailing determinism across the board. The idea is that neither deterministic aspects of our agency nor indeterministic ones are compatible with the tenet that we are morally responsible and free. Not all incompatibilist positions that refute both the libertarian notion of free will and determinism share the notion of agency that Pereboom outlines. The *no-free-will-either-way* position, for instance, maintains that we cannot be free and morally responsible, regardless of whether determinism or indeterminism is true. Galen Strawson defends one version of this view. According to Strawson, it is metaphysically impossible for human beings like us to be free agents. That is why, no matter whether determinism is true or not, we cannot be morally responsible. According to a even stronger version, free will is not just metaphysically impossible, but even conceptually so: the very notion is contradictory.

Contrary to those positions, Pereboom argues that the concept of free agency is both logically and metaphysically well-behaved. Besides, the difference between living in a deterministic world and living in a indeterministic world is relevant for human morality and the rationality of our moral appraisals. Between the two main varieties of libertarian incompatibilism – *leeway* incompatibilism and *causal history* incompatibilism – he thinks that only the second one catches what is essential to freedom and responsibility. Leeway incompatibilism rests on the tenet that having alternative possibilities is not only a necessary condition for freedom and responsibility, but it is a condition with an explanatory import. We are free, morally responsible and thus blameworthy or praiseworthy for our deeds *because* we could have done otherwise than we actually did. Whereas according to causal history incompatibilism, having alternate possibilities as such – even if it turned out to be a necessary condition for being free – cannot explain human freedom and responsibility. An action is free not because the agent could have done otherwise, but because it is not produced by a process that traces back to causal

factors beyond the agent's control, as deterministic processes typically are. In other words, what accounts for an agent being free and morally responsible for her actions is the causal history through which the agent has arrived to those actions. But it is not indeterminism *per se* what makes an agent free. We can think about that in terms of the distinction between open and close future. There is a sense of "open" in which if the future is open, then there is nothing in the present that settles what will be the case tomorrow. Suppose that we live in a universe in which the future is open in that sense: are we thereby free? No, because many things in the future may be unsettled while our decisions being inescapable consequences of factors beyond our control. Thus indeterminism is not a necessary condition for being free.

However, if we share the incompatibilist intuition that if everything is already settled in advance than we cannot be free, we may think that a open future in the indeterminist sense is at least necessary condition for being free. Frankfurt-style examples, though, provide powerful objections to the idea that someone is free and responsible only if she could have done otherwise than she actually did. Suppose that George has to decide whether to lie about his taxes or be honest. Eventually, he decides to act immorally and to lie. However, unbeknownst to him, a neuroscientist has implanted in his brain a device that can detect his intentions. The device is such that were he to form the intention to act honestly with respect to fiscal behavior, it would intervene and make him act immorally instead. In such a situation, George could not have done otherwise than he actually did, and thus the leeway condition for being free is not fulfilled. Yet, we share the intuition that, in the situation described, since the device implanted by the neuroscientist has not intervened, his action has been freely chosen. Contrary to what the leeway incompatibilist theory predicts, he is morally responsible and blameworthy for what he has done even though he could have not done otherwise. If those counterexamples go through, having alternative possibilities is not a necessary condition for being free and the whole project of leeway indeterminism fails. Note that Frankfurt-style counterexamples do not impinge on the requirement of a indeterministic condition on the causal history of the decision of the agent. Indeed, in all such cases, the lack of alternative possibilities is a consequence of "external" factors concerning the situation in which the agent find himself. If we changed the story and assume that the device influenced the causal history in a way to make it deterministic, or in other ways beyond the agent's control, our intuition that an agent is free and morally accountable would fade away.

However, classical Frankfurt-style examples cannot force the libertarian to a causal history version of the position. The intuition that having the capacity of doing otherwise is relevant for our freedom and responsibility is hard to give up. The libertarian can still argue that the intervening device in Frankfurt-style's examples does not touch the agent's capacity of doing otherwise that is essential for moral responsibility. It is true that the agent cannot *act* differently, because if he decided to do so, the device would intervene and prevent him from acting differently. However, an alternative possibility conditions is still at work in explaining why even in those circumstances the agent is free. If the agent could not *decide* to do otherwise, we would not have the intuition that she acted freely and she is morally responsible for her action. As it is sometimes put, the possibility of a "flicker of freedom" is required for moral responsibility. Does this fact confirms that the fundamental idea of leeway incompatibilism, i.e., that possessing alternative possibilities is a fundamental factor in explaining freedom and responsibility? No, Pereboom seems rather to think that if the incompatibilist takes seriously the challenge set by Frankfurt-style cases, a indeterminist condition on the causal history of the agent's actions will emerge. Even if eventually the leeway incompatibilist may be right in claiming that there is an alternative possibilities condition necessary for freedom, that is so only in virtue of the holding of a condition on the causal history of the agent's decision.

Consider the compatibilist objection to the tenet that the possibility that a flicker of freedom occurs is necessary for our responsibility. There are more complex Frankfurt-style cases, in which the device implanted by the neuroscientist is able to detect some previous sign of the agent's decision to act immorally. Thus, if the sign does not manifest itself, the device intervenes and forces the agent to decide to act immorally. Suppose that the sign is blushing at a certain moment t, and that it *does* manifests itself at t. Of course, if indeterminism is true, it is still true that the agent could have not blushed at t, but this flicker is not "robust" enough to ground freedom in the sense required for moral responsibility. Indeed, we have the intuition that whether the agent blushes or not at t is irrelevant for the moral import of her action. It cannot be the occurrence of the blushing or the lack thereof *per se* that accounts for the agent's moral responsibility. The alternative possibilities that can justify the leeway incompatibilist condition on freedom must be such that whether the agent goes for one or the other is relevant for her moral responsibility. If she is blameworthy, then it has to be the case that had she done otherwise, she would

have been praiseworthy and vice versa. But clearly she is not blameworthy for blushing at *t* as such. Whether the agent decide or not to act immorally would be a flicker robust enough to ground moral responsibility, but in the new scenario it is not the case that the agent could have decided otherwise, and thus it is not a flicker at all.

Pereboom sides with the libertarian in the debate over this refined Frankfurt-style cases, since he thinks that although it is not the presence of alternative possibilities as such what makes the agent free and responsible, the refined version of Frankfurt-style arguments still leave the leeway core intuition intact. It is important to notice that the link between the sign and the decision has not to be either deterministic or in any way sufficient for causally determining the agent's decision. If that were the case, then the intuition that the agent is free will be too weak to survive the incompatibilist standards. If the agent is forced to act in a certain way by the occurrence of the sign, then he is not free in any interesting sense. But if the sign is not sufficient to determine that the agent will act in a certain way, than it is still in the power of the agent to do otherwise, and the example does not disproof the alternative possibilities condition on freedom. Indeed, the agent is free only if the possibility of a robust flicker of freedom is still open to her. And the leeway incompatibilist can claim that her moral responsibility is explained precisely by the fact that she is praiseworthy or blameworthy depending on which way she goes.

However, Pereboom parts company with the libertarian because he does think that there are Frankfurt-style cases in which the leeway compatibilism fails. Suppose the sign for deciding to act immorally at *t´* is that the agent at *t* does not consider some strong moral reason to act morally, and suppose that considering such reason is not causally sufficient for her decision to act morally, but only necessary. Furthermore, the neuroscientist has implanted a device that is idle in so far as it does not detect any activity of considering moral reasons at *t*, but it forces the agent to act immorally at *t´* in case it detects moral considerations at *t*. Now, the agent at *t* does not engage in any moral considerations, and consequently the device does not intervene. We have the intuition that the agent is free, even if it is not the case that the agent could have done otherwise, not even in the sense that she could have decided otherwise. Letting apart the details of the discussion, which have given raise to much interest in literature, what is relevant here is to stress that Pereboom version of the Frankfurt-style objection to the leeway incompatibilist is designed to show that the causal history condition is fundamental for

explaining moral responsibility. After all, why in Frankfurt-like cases we have the intuition that the agent is still free and morally responsible? Insofar as the causal history of the decision is not touched by the intervening device or other factors that are not in power of the agent, the intuition of freedom is left untouched by the presence of intervening devices – no matter how subtle and "invasive" they are. Note, for instance, that if we make the connection between the sign and the decision too strict in terms of sufficient causal determination, then we lose the intuition that it is still in control of the agent to do otherwise. That is because if the link is causally determining, then the causal history of the decision would contain aspects that are beyond the agent's control. Therefore, the relevant condition for having freedom and responsibility – the conditions with explanatory power – is having a indeterministic causal history, such that allow for the agent to have control over her decision. If this condition holds, then we can derive some sort of alternative possibilities conditions too – but the core of the notion of freedom that is relevant for moral responsibility does not lie in the presence of alternative possibilities, rather in having control over one's own decisions.

If the condition of alternate possibilities does not catch the core of our notion of freedom and moral responsibility, the problem of determining whether someone "could have done otherwise" is no longer crucial for establishing moral responsibility and freedom. And this is good news because the debate on freedom and moral agency risks to wind up in a stalemate by focusing on the he proper analysis of "could". A compatibilist would argue that a counterfactual analysis of "could" is required in such cases. Very roughly, if there are possible words close enough to ours in which the agent acts otherwise, then it is true in the actual world that she could have done otherwise. But the fact that her choice is causally determined by previous facts beyond her control does not imply that its occurrence is metaphysically necessary, i.e., that there are no possible worlds in which she acts otherwise. Thus, since the agent could have done otherwise, she is free even if she inhabits a deterministic world. The incompatibilist objects to a counterfactual analysis of "could" here, and argue that there is a sense in which if determinism is true, then the agent could have *not* done otherwise, and thus it is not free in a deterministic world. Which sense of "could" should we consider here? If we maintain that the alternative possibility condition is explanatory central for freedom, both are relevant for the question whether the agent is free or not.

And – most importantly – both senses are legitimate, since the nicely match the compatibilist and the incompatibilist supporting intuitions respectively.

The version of incompatibilism that Pereboom puts forward is immune to the risk of finding itself in such a dead-end. According to causal history incompatibilism, we are free only if the causal history of our choices and deliberations involves some essential factor that is under our control. If an agent is morally responsible for her decision to perform a certain action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which she has not control. The point is not only that if an action is a inevitable consequence of what has happened so far in the universe (given the actual laws of nature), then it cannot be a free action and the agent cannot be morally responsible for it (note that leeway incompatibilism, too, may be claimed to catch this aspect). The point is that we are justified to believe that the agent is not free and morally responsible for her actions only in case that her actions have originated from something over which the agent has not control. But being a deterministic consequence of previous events is not the only way in which an event can escape our control. Also a event that happens for no cause at all or randomly may be completely beyond our control.

It is crucial to stress here that indeterminism as such is no warrant of freedom and responsibility – as compatibilists have often stressed. And for the same reason that leeway incompatibilism fails to catch the core of the notion of freedom. Suppose that our world is indeterministic, and more precisely, the processes through which a agent gets to a decision are indeterministic. What we decide will be a consequence of which ones among the alternative possibilities have turned out to be actual, and if we have no control over those events, then we are not free. What, as a matter of fact, is beyond our control can account for our freedom no more if it is a consequence of a indeterministic process than if it is the outcome of a deterministic process. Therefore, no matter whether the causal history of our decisions is deterministic or indeterministic, if there are no crucial elements of it that are under our control, we cannot be free. However, can the processes underpinning our deliberations be such that they are in some relevant way under our control? Libertarians think it can, whereas hard indeterminists maintain that all our decisions are determined causally by things outside our control, since they are *alien determinist events*. Pereboom sides with the libertarian in maintaining that it is

metaphysically possible that a human agent be free because in control of her deliberations. However, he sides with the hard determinist in maintaining that we do not have free will and moral responsibility. His reasons for that claim, however, are not grounded in the truth of determinism: rather, there are very good empirical reasons to believe that we are not in control of any of the events that constitute our decisions.

Actually, with respect to what he calls "event-cause libertarianism" as opposed to "agent-cause libertarianism", Pereboom's position is slightly stronger, since he thinks that event-cause libertarianism encompasses a notion of agent such that, at least by metaphysical necessity, is not in control of her deliberations. I think that the idea of event-cause libertarianism can be made more precise by appealing to the underlying metaphysics of persistence, and in particular, to the distinction between endurantism and perdurantism. According to the perdurantist, agents – as any other entity that persists in time, namely that exists at more than one time – are nothing over and above mereological sums of instantaneous events. Those events are the temporal parts of the agents – those commonly said to be the phases of the agent's life. Within this framework, it is easy to tell what is an agent's decision: it is a temporal part of the agent. What causes a decision, though? If it is the outcome of a deterministic process, it is caused by former parts of the agent in such a way that the agent has no control over the process and thus she is not free. If it is a *truly random event*, viz. something that happens with no cause whatsoever, the agent will not have control over it either. But even if it is a *partially random event*, which the agent cannot causally determine, the agent will not have enough control over her decisions to be free. Therefore, if decisions are events either without a cause or caused by other events, as in the even-based version of libertarianism, it is hard to see how there can be decisions over which the agent has enough control to be free and morally responsible. Since, even the non-random part of the determination can only be another event over which the agent has no control. And the same goes if the libertarian insists that the causally determining factors are things like the agent's character or her capacities. In the event-based version of the theory, what causes an agent to have her actual character cannot be something over which the agent has control.

However, there is a way to add the kind of control required for freedom to the indeterminist picture of the libertarian. In so far as it is coherent to maintain that the agent herself, and not an event (even if one strictly connected

to the agent, as one of her temporal parts), causes her decisions without being determined by factors beyond her control, the notion of a free agent in the sense required for moral responsibility and thus the notion that the libertarian needs to state her position is coherent. Agent-cause libertarianism is precisely the view that it is a primitive feature of the agent to be such a causal source of her actions. Although we can make sense of the idea that the agent as the whole composed by temporal parts, and not any of the parts as such, is the ultimate free cause of her actions, I think that a endurantist metaphysics makes the picture far neater. According to the endurantist, the agent – as any entity that persists in time – persists by being wholly present at each moment of her existence. That is, it is the agent itself, and not any of her temporal part that we find in each phase of her life. Within the framework of a endurantist metaphysics it is clear how the causal relation underpinning the agent's choices looks like: one of the term of the relation is the agent, the other term is an event, namely the choice that the agent has caused to occur (or to whose occurrence the agent has contributed fundamentally).

Now, endurantism is a less revisionary metaphysics than perdurantism, i.e., it is closer to common sense. However, it is also a position less sympathetic to hard sciences. This is true in general, but it is even more apparent in the present case. Modern science makes the notion of an agent as the free cause of her decisions suspicious. The non-reductive materialist strategy to accommodate agent causation within the physical world looks the most attractive, but – as Pereboom convincingly argues for a whole chapter – is not better off than the alternative strategies. At the end of the day, the notion of agent causation of the libertarian violates well-established scientific conceptions. Therefore, even if freedom and moral responsibility are coherent notions, and it is metaphysically possible for a free agent to exist, we are not likely to live in a world inhabited by free agents, and thus we are not justified in seeing us or the other as morally blameworthy or praiseworthy. In other words, the best libertarian version of causal history incompatibilism, namely the agent-based one, has to be abandoned, and the only plausible incompatibilist alternative left is hard incompatibilism. The conclusion is that we are not in control of any of the outcome of our choices, because the causal history of our decisions is entirely made of events over which we do not have control: alien deterministic events, truly random events or partially random events.

Hard incompatibilism has to be defended also from the compatibilist challenge. A compatibilist could agree with Pereboom's picture of a agent as causally determined both in the deterministic and in the indeterministic aspects of the processes underpinning her decisions. Yet the compatibilist ascribes moral responsibility and freedom to humans. According to compatibilists, indeed, causal determination – the sort of lack of control that Pereboom ascribes to agents in ordinary cases – does not exclude free will and grounded ascriptions of responsibility. Therefore, Pereboom has to distinguish his position from its compatibilist counterpart, namely the position embracing both causal determination and moral responsibility, and to defend it as the only viable alternative. To that effect, Pereboom argues that compatibilism fails to spot the relevant similarity between ordinary decisions in which the agent is causally determined by factors behind her control and situations in which the decision is the outcome of a covert manipulation. Since our intuitions in cases of covert manipulations are that the agent is not free and responsible, we should conclude that causal determination in ordinary situations, too, is incompatible with the assumption that the agent is free and responsible. In the present context, I will not discuss the "four-case argument" that Pereboom puts forward to defend his tenet that the two situations are similar in the relevant respect and thus compatibilism fails. Rather, I wish to focus on the modal status of the incompatibilist's notion of determination.

Although Pereboom is right in claiming that the focus on the issue of the proper analysis of "could" leads the debate on free will to a stalemate, I do not think that that is true with respect to all modal considerations about decisions and actions, in particular with respect to the distinction between determination of the future and necessity of the future. Firstly, it is crucial for hard incompatibilism that causal determination does not imply metaphysical necessity. If every human choice is metaphysically necessary, then its occurrence is entailed by the state of the world up to the moment of its occurrence together with the laws of nature (since anything entails a necessary truth), and hard indeterminism collapses on hard determinism. Secondly, the hard incompatibilist's notion of causal determination should not imply nomological necessity either, and essentially for the same reason: if the choices of the agent are determined by what happened in the past together with the laws of nature, indeterminism cannot be true, and full-fledged hard incompatibilism follows the same fate. Can the hard incompatibilist resorts to a notion of "logical" determination, which is weak enough to allow for both

determinist and indeterminist factors to enter the causal history of the agent's decisions? If "logical determination" here means simply that every statement about an agent's choice is bivalent (either true or false, but not both or neither) then I maintain that he can. Maybe there is a stronger notion of indeterminism to the effect that statements concerning our choices are logically undetermined, but physical indeterminism (the tenet that the history of the world up to a given moment together with the laws of physics does not settle all aspects of the future) surely is compatible with the claim that statements concerning our choices are logically determined. Moreover, once it's clear that the hard determinist *should* endorse this notion of determination, because it is the only plausible alternative left, it becomes also clear that in the hard incompatibilist's picture, it is *not* the case that the agent lacks freedom *in virtue* of the outcome of her choices being logically determined. And this is a important difference between hard incompatibilism and hard determinism, which can rest on the idea that we lack moral responsibility because every action of ours is determined in advance. To see the point, consider the following: even in a indeterministic world *with free agents*, the logical sense of determination can be maintained (for instance, if something like "the thin red line view" is true). Therefore, if the agent lacked freedom only in virtue of her choices being logically determined, hard incompatibilist would overgeneralize to libertarianism and it would turn out to be incoherent. That situation forces the hard incompatibilist to put the crucial distinction between the libertarian and himself in some other feature of the causal processes that leads to decisions. For the libertarian there is a causal link between the agent and the choice, which is under the agent's control, whereas for the hard incompatibilist there is no causal link of this sort, because all causal relations that underpin the agent's decisions are relation between event.

In the last three chapters, Pereboom focuses on the moral consequences of his theory, in order to defend it against the charge of making morality impossible. The central idea here is that moral responsibility is only an aspect of morality, and it is not even likely to be the most essential one. Indeed, praise and blame seem to be relevant only to the "irrational" part of morality, and moral value is untouched by them. Actions can be either morally good or wrong, even if they are never praiseworthy or blameworthy. Pereboom does not deny that sometimes a emotional twist may be beneficial for morality, but he argues that for the most crucial aspects of morality there are incompatibilist "surrogates" for praise and blame. In any case, the overall appraisal of the

moral role of rationally ungrounded emotions seems to suggest that we should dispense with them. Pereboom, thus, does not think – as certain hard determinists do – that for pragmatic reasons we should act *as if* we were morally responsible. Living by thinking that we lack freedom and responsibility is morally desirable and within our ordinary capacities. Even if the argument of these chapters probably will not convince anyone who, at this point, is not already both incompatibilist and non-libertarian, they provide lively challenges for libertarians and complete the overall plausibility of the hard incompatibilist view of reality.

Interview

Sean Spence

Edited by Duccio Manetti*

We very much regret to inform that Professor Sean Spence died on Christmas day (2010) after a long illness. This interview is probably one of the last expressions of Professor Spence's thought. We are grateful to him for being so kind and helpful as to honor us with his exhaustive answers despite his illness.

SEAN SPENCE was Professor of General Adult Psychiatry at the University of Sheffield. Psychiatry and Philosophy has lost a great scholar. Our sincere and heartfelt condolences go out to his family and loved ones.

1. Your latest book is entitled *The Actor's Brain*. Can you tell us what is an *actor's brain*?

The purpose of using the term the 'actor's brain' was to identify what I hoped would be captured and characterized over the course of the text: namely, those conditions (anatomical, physiological, psychological, etc.) which must pertain within the nervous system of a human being in order for them to be seen to be performing, what appear to be, 'purposeful' acts in the world. In other words: What is it within the systems of the brain that 'supports' the emergence of apparently voluntary behaviour? Such an account could not be exhaustive (hence, the book's subtitle: '*Exploring* the Cognitive Neuroscience of Free Will'). However, I was very concerned that it should be grounded in neurobiology, unapologetically building upon what is known of neural function, while also eventually arriving at behavioural, phenomenological distinctions that would be recognizable to a philosophical readership, e.g., the difference between 'actions' and 'movements', between 'purposeful' behaviours and mechanical 'events', as these might be understood by an author such as MacMurray (1991). One of the most prominent themes to emerge across the book was that of *constraint*: the limitations set upon the extent of

* University of Florence

our freedom and manifest within many domains (e.g., in our neuroanatomy, neurochemistry, and indeed our subjective temporal awareness, as revealed by Libet and others).

I was also mindful that much of the second half of the book would deal with 'real-life' situations, encountered in the clinical arena, where voluntary behaviour is either mechanically aberrant or undesirable in terms of its valence; situations in which the constraints upon the human actor become even more obvious, forcing the clinician/observer to consider factors ranging from what one might call our 'brute neurology' to rather more diffuse interpersonal, social influences. This would lead on to the raising of some pivotal questions concerning the future of psychotherapeutics (Chapter 10), namely: Whether, once damaged, an actor's brain may be restored to volitional function? Might freedom return?

2. The acts performed by the subjects in Libet's experiments seem too simple and not enough representative of everyday life decisions. Do you think it is possible to include value choices in the experimental sets?

I think the experiments performed by Benjamin Libet and colleagues in the 1980s were necessarily simple in some aspects of their design, since they sought to strip a voluntary act to its minimal constituents: a single subject introspecting about their motor intentions, while they made self-paced movements of their right index finger or wrist. By acquiring objective electroencephalographic (EEG) and electromyographic (EMG) response data, together with the subject's internal estimation of the time of onset of their own 'intention to act', Libet et al (1983) were able to elicit their central finding: namely, that the EEG antecedents of a voluntary act arise in the brain *before* the subject's *conscious* intention to perform that act. The simplicity of this design is part of its beauty.

Now, there is a vast array of more complicated acts that one might wish to study (as opposed to the simple movements of a finger). However, I think this is essentially a question of empirical ingenuity: designing experiments that may sequentially access actions of increasing complexity, e.g., the learning of motor skills, the generation of novel behaviours, the formulation and expression of moral preferences, the telling of lies, etc. This is a theme that I return to throughout the book.

3. Do you think that a neuro-philosophy, as Henrik Walter pointed out, could be useful for a science of volition?

Yes, I think there are certain areas where the interests of neurology and philosophy overlap sufficiently for synergy to emerge. Such areas of convergence may also yield useful insights into disturbances of volition (Spence 1999).

One area that provides a clear example is the problem of hysteria (or conversion disorder), which I address in Chapter 7 of *The Actor's Brain*. As the reader may be aware, the sort of problem we encounter in hysteria may be such that a patient presents as being unable to move her arm, for no apparent reason. Medical examinations and investigations are 'negative' and it seems to the doctor that there is no biological, physical explanation for the symptom (i.e., paralysis). In addition, the impairment appears to come and go; it may be present in company but not when the patient believes herself to be alone, unobserved. The patient says she cannot move yet, medically, there is no apparent impediment to her movement. Eventually, the medical 'explanation' offered is that there is some unconscious process that prevents the patient from moving (a process which serves to resolve a latent conflict of some sort). To borrow one of Freud's examples, it may be that a young woman exhibits a paralysed right arm, which impedes and (thereby) conceals her unconscious desire to hit her father (Freud and Breuer 1991, pp. 93-94). Somewhat anachronistically, this form of hysteria has served to enshrine a Freudian understanding of the mind in the psychiatric diagnostic systems currently available to us (e.g., the DSM IV; APA, 1994). These systems each require the physician to diagnose hysteria/conversion on the basis of the patient exhibiting a functional deficit, which is neither attributable to a physical cause nor (explicitly) the product of feigning (i.e., malingering). However, this distinction between hysteria (unconscious, unknowing causation) and malingering (conscious, knowing causation) is impossible to justify on empirical grounds; unless, that is, one believes that the physician can tell what the patient is thinking (Spence 1999)!

However, there is another way of formulating the problem of hysterical conversion, which, informed by the language of action philosophy (and cognitive neuropsychology), actually points us towards the likely causal mechanisms at play. For, while the Freudian formulation of our patient's paralysis emphasizes her inability to move due to the influence of some

unconscious force (outside her awareness), closer examination of hysterical phenomenology in the light of action philosophy suggests something quite different (Spence 1999). The patient exhibiting a hysterical paralysis maintains her symptom while she is awake and alert; she 'loses' her symptom when sedated or distracted. Indeed, it is this symptomatic inconsistency that prompts the diagnosis (above; though notice, that the same phenomenology would arise in malingering). Hence, it is likely that the patient's attention to action is, in some way, necessary for the maintenance of her paralysis; it is not an unconscious process. Therefore, while we may draw a philosophical distinction between an 'action', chosen by an agent, and a motor event, or movement (e.g., a reflex), arising automatically, what we appear to have in the hysterical symptom is an example of the former action, a voluntary action (no matter how aberrant): the patient's attention to action is pivotally implicated in its maintenance (this is the opposite of what we might expect were the symptom to be maintained by a Freudian unconscious).

Furthermore, if we then go back to the empirical literature, we find that certain objective (e.g., EEG, EMG and ergonomic) measures acquired from hysteria patients *do* in fact support the contribution of so-called 'higher centres' to the 'performance' of hysterical symptoms (see Chapter 7 of *The Actor's Brain*). Hence, combining an analysis of hysterical phenomenology, with the vocabulary of action philosophy, and the acquisition of more subtle biological measures leads us to a deeper (and contrasting) view of the nature of hysteria: it is not a product of unconscious desires but may be understood cognitively as the product of an executive system (where conscious awareness assists in its maintenance).

4. If we discover an abnormal situation in the volitional processes like one of the patients with anarchic hand syndrome, can we infer that a 'normal' agent exists and acts somewhere in the brain?

In the case of the anarchic hand syndrome, where a man may experience his right hand as reaching for and grasping objects inappropriately, 'against his will', I think what we are witnessing is evidence that agency may be frustrated. It is as if an automatic sequence of behaviours (a 'schema', in the vocabulary of Shallice 1988) had been liberated from the hierarchical control of the motor system as a whole. Hence, the limb appears to behave autonomously: the man's agency does not encompass his affected limb. He retains awareness of the

discrepancy and this suggests that 'somewhere' within the nervous system there is a rational actor 'looking on'; he cannot exert control over the limb but he knows enough of his plans as to know that they are not being 'obeyed'. In Chapter 5 of the book I deal with the different forms of anarchic and alien limb that may arise, and what seems common to them is that the patient, the frustrated agent, retains an awareness of what they would like their limb to do or refrain from doing, yet they cannot control it. Hence, they continue to experience themselves as agents (with preferred goals), but they are faulty agents, agents who cannot realize those goals.

Indeed, we might contrast such patients with those who experience what seems to be an even more profound disturbance of agency: namely, utilization syndrome. For while the anarchic hand patient knows that their limb 'will not do what I want it to do', the patient exhibiting (severe) utilization appears not to notice that their limbs are interacting automatically with the environment. Hence, if a pen is left on the table they will start to write with it, if there is a cup they will drink from it. They may even perform quite complex behaviours, in response to environmental cues, apparently without any awareness that their behaviour is being manipulated.[1] So, in this case, we seem to witness both the disturbance of objective movement (control of motor events, for these are not chosen 'acts') and the absence of a subjective agent (since, in extreme cases, the patient/subject seems unaware of their lack of volitional control, their manipulation by their surroundings).

5. Some philosophers, like Dennett for example, consider Libet's experiments too Cartesian. Libet's original intention was to discover and legitimate the mind against or beyond the brain. Do we have to reformulate these experiments? Are they corrupted by a mild form of dualism?

I think it is inevitable that Libet's experiments be conceptualised in dualist terms, merely because of the methodology used and the questions he asked. In essence, he was examining the correlation between certain subjective first-person phenomena (the perception of an urge to move) and externally detected (objective) third-person phenomena (EEG and EMG signals, the latter indicative of movement). So, his results would inevitably consist of a temporal comparison between the emergence of a highly subjective event occurring in

[1] See Lhermitte 1983.

'inner space' (the intention to move) and a verifiable, manifest event arising in the outer world (the movement itself). Hence, to adequately understand his findings would seem to require a solution to the 'hard problem' of consciousness, although further empirical refinement would still be necessary to distinguish correlation from causation.

6. In *The Actor's Brain* you say that neuroscience is searching for the 'it': do you think that this 'it' could be the intentions? Which role do intentions play in the volitional mechanism?

When I mention an 'it' I am really attempting to describe the *source* of intentions, whatever it is that precedes our conscious choice, both in terms of its temporal and neurophysiological characteristics.

7. One of the purposes of your book is to deconstruct Libet's arguments. Can you explain how this is possible?

As I state at the beginning of the book, I regard Libet as having made a major contribution to this field and it is because of its importance that I seek to clarify what it means. One way of summarizing his contribution is to say that he demonstrated the *temporal* constraints impacting human volition: whether that is our awareness of our own agency (becoming aware of our intentions only after they appear to have been set in motion) or our awareness of 'incoming' sensory data (only becoming aware of sensorimotor phenomena (qualia) after a finite period of specific neurological activity, so-called 'neuronal adequacy', has taken place).[2] Hence, what several of his experimental designs serve to show us is the limited extent of our agency. If I only become aware of intentions to act after their related act has begun to emerge from the brain then to what extent am I in control (Spence, 1996)? It strikes me that Libet's work emphasizes volitional constraint and there is an account that may be given of the many constraints that impact our apparent volitional freedom (and I deal with these in ensuing chapters of the book: temporal, neurochemical, socially hierarchical, etc.).

But it is also possible to critique some of the conclusions Libet derived from his own work. For instance, he argued that free will was still justifiable if it

[2] See Libet 2004 for an overview.

functioned as a form of 'veto', a kind of 'free won't' active prior to actions. Hence, being aware that an emerging action was inappropriate the subject/agent could decide to stop it or change course. This would provide evidence of freedom. My response to this is that if Libet's basic findings are correct, i.e., if a period of neuronal adequacy is necessary for us to be aware of subjective phenomena (including our own thoughts), then the veto thought, the idea of stopping an ongoing action is itself likely to be the product of foregoing neural activity (arising out of awareness). So the veto thought is just as 'post hoc' as the initial 'urge to move'. They both appear to arise in subjectivity after neural control mechanisms have commenced. So, if Libet's neuronal adequacy hypothesis is correct then, the veto does not preserve the libertarian's notion of free will.

8. Do you think agency is an important topic in the investigations about free will?

Yes!

9. Is it correct to think that free will is an evolutionary instrument that biology gives to humans in order for them to direct their own behaviour?

Clearly, this would be a highly teleological way of understanding the outcome of evolution. What seems to be the case is that the existence and optimal functioning of the human nervous system supports the generation of what appear to be purposeful behaviours under certain circumstances. Nevertheless, as we examine each of the many domains of biological, psychological and social influence at play within and around us, we find that we can increasingly identify tangible contributors to human actions, or at least, apparent constraints upon its parameters. This has led me to focus on the idea of a 'Human response space' (the subject of Chapter 10 in the book). What I am trying to get at here is the idea that there might exist a finite, though probably highly variable capacity for freedom, varying both between and within individuals over time. Hence, we might be each capable of acting freely under optimal conditions but the opportunities for those conditions to arise and the specifications of 'optimality' might vary greatly between individuals. The man who sits in the refugee camp, close to the point of starvation may exhibit less purposeful behaviour than the choreographer in prime physical health who is at the height

of their powers. These are dramatic examples but the book abounds with more subtle examples: e.g., the extent to which prefrontal lobe dopamine metabolism may impact inappropriate repetitive behaviours, or serotonergic dysfunction relate to violent self-harm, or the presence of an apparent authority figure sanction the performance of cruel acts towards a stranger. There are many domains of influence that may distort or constrain Human response space.

10. Some philosophers think that the real problem of free will is to define exactly what this very concept means; for example if it corresponds to the intentions, or the power of acting or to long-term decisions and choices. Do you think philosophy could help neuroscience clarify the notions in this field?

Help in fine analysis of action and avoidance of sloppy thinking (remember Libet paradox, veto and my long-term comments...)

11. Can the movements that a player makes in sports like basketball or soccer be considered an example of the gap between automatic acts and conscious deliberative acts?

Yes, absolutely, Shallice – schema, increasing automation with practice.

12. Your recent work is about the neural correlates of deception: does it invoke the function of 'higher' brain systems?

Chapter 8, summarizes. Of interest to a neuroscientist not least because it is one of those areas where imaging may inform us of something we did not know already. As in the case of hysteria, where we wish to make a distinction that cannot be justified empirically in the clinic (i.e., between 'hysteria' and feigning), here we have the distinction between truth and lying, a distinction that most humans can judge at little above the level of chance (Bond and De Paulo 2006). Furthermore, it is another example of an executive control process, one that a subject must attempt to deploy in real-time, e.g., when calling to mind, suppressing or creating new scenarios.

REFERENCES

American Psychiatric Association (1994). *Diagnostic Criteria from DSM IV*. Washington DC: American Psychiatric Association.

Bond, C. F., & De Paulo, B. M. (2006). Accuracy of deception judgements. *Personality and Social Psychology Review, 10*(3), 214-234.

Freud, S., & Breuer, J. (1991). *Studies on Hysteria*. (Tr. by J. Strachey & A. Strachey). London: Penguin. [First published in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. II. London: Hogarth Press and Institute of Psycho-Analysis, 1955]

Lhermitte, F. (1983). 'Utilization behaviour' and its relation to lesions of the frontal lobes. *Brain; 106*, 237-255.

Libet, B. (2004). *Mind Time: The Temporal Factor in Consciousness*. Cambridge, MA: Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intentions to act in relation to onset of cerebral activity. *Brain, 106*, 623-642.

MacMurray, J. (1991). *The Self as Agent*. London: Faber and Faber. [1957]

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Spence, S. A. (1996). Free will in the light of neuropsychiatry. *Philosophy, Psychiatry & Psychology, 3*(2), 75-90.

Spence, S. A. (1999). Hysterical paralyses as disorders of action. *Cognitive Neuropsychiatry, 4*(3), 203-226.

Interview

Daniel Dennett

Edited by Marco Fenici* and Stefano Di Piazza*

DANIEL CLEMENT DENNETT (1942, Boston, MA) is an American philosopher and cognitive scientist. His research work mostly concerns the philosophy of mind, the philosophy of science and the philosophy of biology. He has been one of W. V. O. Quine's students at Harvard University, where he graduated in philosophy in 1963. In 1965, he achieved his Ph.D in philosophy at the University of Oxford under Gilbert Ryle's supervision. In his academic career, he has taught at the universities of Irvine, Harvard, Pittsburgh, Oxford, and the École Normale Supérieure in Paris. Since 1971, he is professor of philosophy at Tufts University, where he co-directs with Ray Jackendoff the Center for Cognitive Studies. Among his books, *Content and Consciousness* (1969), *Brainstorms* (1978), *The Mind's I* (with D. R. Hofstadter, 1981), *Elbow Room* (1984), *The Intentional Stance* (1987), *Consciousness Explained* (1991), *Darwin's Dangerous Idea* (1995), *Kinds of Minds* (1996), *Brainchildren* (1998), *Freedom Evolves* (2003), *Sweet Dreams* (2005), *Breaking the Spell* (2006), and *Science and Religion: Are They Compatible?* (2011).

1. Professor Dennett, thank you very much for accepting this interview with *Humana.Mente*. Consistent with the topic of this issue, I would like to discuss your *Freedom Evolves* (2003). Can you explain to our readers what urged you to write a book about the problem of free will?

Let me start by saying that we should anchor the concept of free will to the fact that people think that it is important. I can define free will in such a way that we do not have it but this is little interesting. Instead, there is a variety of concepts of free will worth wanting that we are talking about. What is important about free will is that it gives us the chance to be moral agents. Chimpanzees do not have minds that can appreciate what murder is. Accordingly, if a chimpanzee

* University of Siena

kills a person, this is not murder. Similarly, we do not hold children responsible or retarded people, in that we do not think they have free will in an important sense. Instead, it is normal adults who are morally competent. It is to this capacity to be moral agents, therefore, to be moved by reasons, to be able to preserve and protect our mental autonomy through time – as when we discover, for instance, that other agents are intending to usurp our autonomy or to manipulate us – that the concept of free will in which I am interested is connected.

2. A long philosophical tradition denies that we are free because of the supposed truth of determinism. Indeed, many philosophers have argued, we cannot say that we are free to act if what is happening in a second is already defined by the current situation of the world. How can you develop a concept of free will that is important to moral agency while at the same time it does not oppose determinism?

The key concept, herein, is our ability for *anticipation*. We are free to act because we are able to look at the world that we are in and anticipate likely futures, evaluate them, and then avoid the ones that we evaluate as less valuable. It is this capacity that makes us moral agents. At the same time, this capacity does not define a variety of free will outside of the deterministic natural order. A simple case of that is if I throw a rock at your head, and you duck, avoiding getting hit with the rock. Thanks to the deterministic path of light rays, being regular and predictable, the light that bounced off the rock into your eyes allowed you to anticipate the trajectory of that rock with great accuracy. Thus, determinism is actually our friend, because it provides regularities in the world that we can exploit. It is this mega-capacity to secure the good and to avoid the bad that is the essence of free will. You want to have your will cause you to move in the best directions by your assessment of the world's situation. You have desires and intentions you would like to fulfil. Perception causes you to acquire facts about the world that are relevant to those desires. When all goes well, those facts about the world cause your body to make the choices that will most probably satisfy your deepest intentions. That is what free will worth wanting is: the capacity to be guided to effective choices.

3. Your concept of free will is thus preferentially related to the complexity of the relation between human agency and the environment. We act freely because we are able to cope with unexpected changes in the environment. However, the concept of free will is traditionally connected to the perception of our own agency. We retain us to be free because we feel free when making our choices. Do you think that this perception of free will is an illusion?

No, that is not an illusion or, if it is, it is like the user's illusion on our computer. When using our computer, we have the sense that we can open and close files, and we can move them around, drag and drop, and so forth. The actual processes involved behind the scenes in the computer are more complicated than the icons suggest: the interface between the user and the computer is a valuable – because vivid and memorable – simplification of the actual events. Our brain has a lot of user illusions as well – which is a good thing. They help us coping with all the complexities of our brain by oversimplifying the vision of what is going on in it. According to this user illusion, we are not wrong when we see our future as open. In fact, our future is open in a very important sense. I will take a deliberately simplified example. If you play chess against your computer, your computer has the perspective that the future is open. Indeed, it has the sense that it can choose any of the legal moves and that so can you. If that presumption is built into the control of the software, then your computer – although it is an entirely deterministic system – has the perspective that the future is open. But it has a perspective of openness because that perspective is required for making choices. Similarly, our perception to be able to act freely is related to our perception that the future is open – that is, that we perceive ourselves as capable of choosing between different real alternatives. This does not means that the future actually is open, but it is important to make our choices.

4. Thus, your compatibilist view acknowledges that determinism is true while at the same time this does not affect the reality of free will. However, many have claimed that free will is opposed to the truth of determinism. For instance, many neuroscientists nowadays reject human free will based on Libet's (1985) findings, which correlate neural events and the phenomenological experience to decide to act. According to them, Libet

demonstrated that our brain decides to act before we do. What do you think about that?

I think that the interpretation of Libet's findings as demonstrating that specific individuated neural events are the real causes of our decisions to act is the huge artefact of a mistaken conception of consciousness. Libet's work is a perfect paradigm of how you get in trouble if you are what I called a "Cartesian materialist". If you think of consciousness as being something that happens in one place of the brain, to which all contents must be moved to be experienced, then you make a big mistake. But look at how you put the questions: "the brain decides to act before we do". Who is this *we*? Where is it? If you are thinking of the *we* as being somehow resident in one place or another somewhere in the brain, and you are thinking that the decision is already made before it arrives there, you are just making a huge mistake. When I wrote *Consciousness Explained* (1991), I was specifically trying to expose this error. I used Libet's work as my chief target. Amazingly, I underestimated how potent the seductiveness of Cartesian materialism is. People just will not give it up. They all agree when I say that Cartesian materialism is a bad idea, and then two minutes later they are right back using conceptualizations which would only make sense if they are supposing the Cartesian theory. However, if you do not permit yourself the mistake of the Cartesian theory, you cannot formulate Libet's results in a way that looks problematical at all.

5. The interpretation of Libet's experiments is significant not only to the very issue of free will. Indeed, many have claimed that, if our free will does not determine our behaviour, then our social concepts should be changed. For instance, if free will is an illusion, then a retributivist conception of justice, according to which criminals are judged and punished based on the harm they have caused, is not acceptable anymore. They would not be morally responsible for the harm they had caused.

That is what Greene and Cohen (2004) say. There are two things to say about this. First, what Greene and Cohen claim is subordinated to the condition "if free will is an illusion", which is not the case. But, in any way, one of the main points I was trying to make is that the relationship between a mild gentle human retributivism and what we know about the brain is intact. I resist vigorously the idea that we should abandon all elements of retributivism from

our view of the law, and move to a sort of pure medicalization idea. Because, if we did that, we would no longer be able to make the distinctions we need if we are to apply the law, not just to punish the guilty criminal but also to enforce the signing of contracts, for instance. You need the concept of a morally competent agent for that.

6. By "pure medicalization idea" you perhaps mean a consequentialist view of justice, according to which judgement and punishment are inflicted according to the social benefit of their consequences. Why, according to you, staying away from consequentialism is so important?

Because it does not permit the protection of freedoms. Consequentialism treats law violation as a sickness; and, if you're a sick person, you go to an institution to cure you. In an authoritarian state, when people say they have broken the law because the law is unjust, they get the reply: "Well, your brain is not sufficiently mature for that, and we need to cure you by appropriate punishment with good social consequences". Well, we want to be able to restore ourselves to freedom by taking the punishment because we broke the law. If you get a car speeding you would pay a ticket, you do not want to be sent to the speeding hospital for a month. And of course if you really get rid of retributivism and if your model is a medicalization model, then one cannot distinguish the morally competent, adult agents from those who are not able to make promises, sign contracts, and so forth.

7. If consequentialism cannot grant the conception of human being as morally responsible agents, how do you think that a retributivist conception of justice may do better? After all, even retributivism has its flaws. "An eye for an eye, a tooth for a tooth" is not what we may call a valuable saying for a modern conception of justice. In some cases people are not morally responsible for what they do (e.g., children, mentally disturbed people), thereby the punishment nor addresses an intention to break the law neither has a positive effect for the society. Considering these cases, how can a retributivist conception of justice be preferable to consequentialism?

There's a process that protects us all from the excesses of retributivism. We legislate past laws, then, in order to keep respect of the law, we acknowledge very exclusive conditions not to apply it. This creates the opportunity for

people to try finding loophole, trying to exploit the exclusive conditions we have introduced – everybody always tries to exploit any law, whether it's a tax law, or any kind of law. Thus, we have to come back legislating, and to do something to prevent people from exploiting the special exclusive conditions we previously introduced. This creates an "arms race" between exemptions and exploitation of exemptions that is the source of stability in the law. I would like to take a very simple case: how old do you have to be in Italy to have a driving licence?

18 years old.

Ok, it is 16 in the United States. Is one of them right and one of them wrong? No, and maybe you can make the case that you have to be a politician, not a scientist, to lower the age to 16 in Italy – or to raise the age to 18 in the United States. In any case, once we have made that decision – for instance, we decided, that you need to be 18 to drive a car – we fix a legal threshold. If you are not 18, we do not look never more at how mature you are, we do not care if you are a genius, or the most literate person in the world, if you are not 18, then you cannot drive yet. On the other side, if you are 19, then you can be pretty idiotic, you can be pretty dangerous, and you still have the right to drive. Then we adjust that in turn. Indeed, we say: "Maybe, if you are caught during this or that you are going to lose your licence for a while, or maybe you cannot drive at night, or a truck, and so forth". All of this is done to provide some bright lines as the law says where Nature has not done big bright lines. It is not that when you wake up on your eighteenth birthday your brain changes so we can expect from neuroscience another source of evidence about where to draw these bright lines (e.g., about the right age to drive). There is nothing to say exactly. At what age people should be allowed to drive is a political issue, and we want to keep it that way. We want to be able to rely on the world's future, and do not allow anyone to deprive our liberty or our opportunities to set these edges because somebody has decided that we are not competent anymore.

8. Resuming, you are suggesting that we are always judged in front of the law according to a standard of competence which is not fixed by Nature. Accordingly, we do not want to delegate our capacity to fix these standards. Any reasonable view about justice should keep retributivism – i.e., the principle that we are judged according to what we did – but also the

principle that the standards defining the moral competence according to which we are judged may change. Is this correct?

Exactly. That is why I think that a mild form of "revisable" retributivism is preferable.

9. We long discussed the incompatibilist position of the eliminative materialist. Considering another incompatibilist position, the libertarian, who claims that we do have free will not just because we are able to protect our mental economy but also because we are able to do that *by ourselves* – that is, without being caused by anything external to us – may object that your definition of free will is very weak. What would you reply?

Yes, my concept of free will would seem like a weak one to the libertarians. To act freely, they claim, you have to act independently from external causes, that is, indeterministically. However, what they have not shown is why indeterminism would make it any better. I have argued that, if what I do is completely random, then I am no more responsible for that than if it is determined: deterministic chaos essentially is indistinguishable from randomness. I think it should be an embarrassment to the libertarians that the very models we have of randomness is throwing the dice, or flipping a coin. Indeed, those models are chaotic and important, but they are not random. Instead, I argue, those models are useful, and they indicate a way to construct a concept of free will that is relevant to the definition of us as moral agents. Indeed, the unpredictability of chaotic events – such as flipping a coin, or something like that – is the kind of unpredictability needed to decouple from features of the world, hence to act freely. We show to have this kind of unpredictability in our behaviour, but actually even animals employ. The rabbit that runs from the fox takes a very chaotic trajectory. The butterfly moves very chaotically. These are evolutionary adaptations that make these animals harder to catch. That kind of randomness, that kind of freedom is all around us in Nature. It is not the kind of free will the libertarian would like to have for the human species. It is not free will, either. But it is all we need in order to construct a concept of free will that is relevant to the definition of us as moral agents.

10. You seems suggesting that free will is not construed as a sharp-bounded concept. This poses the issue of how free will is obtained. From an evolutionary standpoint, should we say that free will has been gradually construed during the history of life on Earth?

Definitively yes. From an evolutionary standpoint, the fact that only one species currently has free will is only an empirical fact. Maybe in the future there will be more. The important thing is that free will is a new phenomenon in the biosphere. That means that free will has nothing to do with the physics, indeed the physics has not changed since the origin of the Earth. What has changed is biology. There has been an explosion of evitability in our world. The earliest forms of life could not produce any significant behaviour to chance their destiny. Consider again my examples of the rabbit or of the butterfly that move very chaotically. Those are all examples of avoiding activities. However, our nervous systems have more estimable competences that animals and even our ancestors did not have. We can avoid all kinds of things, and of course we can even avoid avoiding, and we can avoid avoiding avoiding, and so forth. We have all recursive capacities to avoid things that we can anticipate.

11. The absence of sharp boundaries to the concept of free will also poses an issue with respect to the ontogenesis. Should we say that children are already born with the capacity to act freely, or is it acquired during child development?

I think both. Even small babies have the fundamental capacity to address the world and to make simple choices – e.g., whether to lift this or that hand up. However, at the beginning of their lives, they have not yet coordinated their sense of action, and, with this respect, they still do not act freely. This is something which requires more time. We know they go through a period when they are simply unable to avert their eyes from a stimulus, and the capacity to move your attention away from one object to another is actually something that requires maturity. Until you do not have that, I think you do not have much of the bases for free will.

12. Herein, I see an important issue for your proposal. It seems to me that you believe that language is the most important ability we need to detach form contextual stimulation, thereby even to become free agents.

Yes, language certainly is very important. In order for an agent to be moved by reasons as concept, it is very important for her to have language.

13. Now, I would like to understand the reasons why you think that language is really important. Let me now just quote your *Kinds of Minds* (1996, pp. 146-147): «Of all the mind tools we acquire in the course of furnishing our brains from the stockpiles of culture, none are more important, of course, than words – first spoken, then written. Words make us more intelligent by making cognition easier, in the same way (many times multiplied) that beacons and landmarks make navigation in the world easier for simple creatures». You are suggesting that language scaffolds though in that word learning increases our cognitive capacities. This proposal alone raises some perplexity to me. Word learning does not seem a fundamental ability to language development. In fact, even animals can learn words by associative processes. Instead, other features of language really might make the difference to the development of our cognitive abilities. For instance, we know that children exploit syntactic hints in their word learning processes...

Well, animals do not really learn words. They learn sounds that have associations, and that is a big difference. I think Terrence Deacon has important things to say about this in his book of *The Symbolic Species* (Deacon 1997), and also my colleagues Ray Jackendoff in his book *Foundations of Language* (Jackendoff 2002). What I really like about Ray's recent work is that syntax is still important but it is no longer the driving machine of language acquisition. Syntax is a feature of words, but words are a sort of semi-autonomous entities, which appropriately move from one language to another. I have been recently thinking of words as a sort of Java Applet. On your laptop, you have a Java Virtual Machine, which permits people to write Java applets who will run beautifully on your laptop no matter what the architecture of the laptop is. Similarly, you have a sort of EVM, an English Virtual Machine. That permits me – without I have to know how your brain works – to talk to you and to know that the words I am telling to you play roughly the role that I intended to play because you have the EVM system for realising those words when they come in. This is the reason why words are not sounds. Sounds are just means that can be pronounced; however, these means convey informational structures, they are like software to the brain.

14. If syntax is secondary relevant to develop language, what is distinctive of words that can alone provide the complexity of language?

I think that words open up the explosion of cultural transmission because they are the key to the digitalization of language. When I talk about digitalization, I am thinking to the fact that, for instance, when you download something from the web, there are lots of tiny variations in the voltages but the finally digitalized value is either 0 or 1. In the end, every voltage is corrected to a pre-fixed value by a norm. The same thing is true for the words. Digitalisation gives language its fidelity. This is very important. You cannot transmit anything without a set of basic fixed elements. This is clear if we look at primates. Consider, for instance, chimpanzees. They exhibit a smattering of culture but they cannot do anything combinatorial – certainly, they can pass along few local techniques for breaking up a nut, for instance, or for fishing termites, but really they cannot put them together in interesting ways. And so their capacities for transmission are very little.

15. I see. Therefore, the ability to learn words, and not just sounds, is what you think grounds our higher level cognitive abilities. Is language equally important to acquire the ability to enter what you called the «intentional stance» (Dennett 1987)? This is a question that always bothered me. Indeed, the capacity to assume the intentional stance is connected to the capacity to attribute rationality, and rationality is clearly a normative, social concept. However, the intentional stance might also be hard-wired in our brain as a case limit of the design stance. On which side of the Nature/Nurture divide should we put the capacity to enter the intentional stance?

On both sides. I think that, at its bases, the capacity to enter the intentional stance is like an instinct, that in principle we might share even with animals. And indeed, there have been a lot of research in the last 35 years on the so called theory of mind that shows that animals do attribute cognitive states to others – at least to some degree – and that certainly human beings, even from the very young edge, are already alert to picking up the symptoms of the intentional system. Despite of that, there is something in the capacity to enter the intentional stance that is the outcome of cultural inheritance. In fact, very often we over-attribute understanding and rationality to animals and to young

children. In a sense, we deliberately do this, we treat children as more rational than they are, and this behaviour provides them some scaffoldings. Due to our tendency to treat children as more rational than they are, children grow a more mature capacity to adopt the intentional stance.

16. Before finishing this interview, I would like to discuss more in general the extent of your research work. In listening to you giving a speech some days ago, I felt as if you have some sort of social, or ethic, aspiration in your work. You are concerned with the problem of free will because it is related to people's concept of human agency, and you want to change people's way of thinking about moral agency. Is this correct?

Yes, it is. I realised that I am opposing a tradition that is several thousand years old but it is simply a mistake to think that free will in the morally important sense is in any conflict with determinism. Now, various people have realised that over the millennia. I think that early appreciations of this were not very convincing too many people because we did not have conceptual tools to take carefully about reason and intention, but now we do. I think that the idea that moral agency depends on physics, or on the indeterminism of physics, is not just a mistake, but a sort of crippling mistakes. It is a confusion that can lead to seriously pernicious social consequences. For instance, Vohs and Schooler (2008) showed that people who read a passage explaining that they do not have free will are more likely to cheat. I think that, if that vision really takes hold, this can be a misconstrual of the science, a one that is really socially unfortunate.

17. So we have that bad philosophy made from scientists might bring people to have bad ideas about moral agency?

Yes, and I think that it is philosopher's personal responsibility helping everyday folk to understand the implication of science. Right now there is a lot of confusion on this very score. I think that scientists are very good at confusing things, and who better should do clarification work than the philosophers? So we have a job. It is an important job.

18. In conclusion, you think the idea that free will is not compatible with determinism is a false myth, which should be abandoned. I wonder

whether you think that philosophical analysis should bring people to abandon other concepts – specifically, the concept of God. I would like to ask you for a comment on an Italian contemporary debate. Giulio Giorello recently wrote a book, *Senza Dio. Del buon uso dell'ateismo* (Giorello 2010), in which he supports the importance of atheism as a value for democratic societies. He argues that denominational dialogue is not enough to democratic societies if they do not also respect the opinion of those who are not the followers of any religion. Such an idea is open to two interpretations. On a weaker politically-correct reading, religions should be ready to confront even with those accounts rejecting the existence of God. Accordingly, atheism represents a social value because it sticks the public debate on values to a human dimension. On a strongest impolite reading, which maybe Giorello is also supporting, religions are supposed not to be able to dialogue with anyone denying the existence of God. Therefore, atheism is a value because it remembers the importance of reason against any form of absolutist obscurantism of reason. On this second reading, the concept of God becomes more harmful than neutral. Which kind of reading do you think is the more proper to the current situation?

I think that this is a delicate political question, not a metaphysical issue. What we are currently seeing is the continuous retreat of religious conceptions of the world in the face of the advance of the scientific understanding. This is a painful process, and we should recognize that a lot of people had a lot of trouble with it. We should accept that well meaning and intelligent people are trying to devise gentle revisions that will preserve as much as possible of their traditions. I think that is a respectable attempt in what I think it is just postponing the inevitable. We should be firm and as polite as we can be with it, but we should not continue to honour the invocation of mysticism and irrationality, and treat it as if it had some privileged position in the space of public reason. I think that the impolite atheists would say – as I myself used to say – that other religions very often want to "play intellectual tennis without a net". They use reason when they think they can score points, and, as soon as they are going get stuffed, they play the faith card, and they switch to a different game. I am simply not going to play that game anymore.

19. So you are more on Giorello's side... but what about the concept of God? Do you think we should get out the idea of it, or is there a chance to keep it? Let me translate the question in other words. Feuerbach thought that religion is anthropologically grounded, thereby, it cannot be eradicated from the image the Man has of himself. On the contrary, Marx thought that religion is a part of the Super-structure, thereby a day we will be able to get rid of it. In the debate between the two, you seem to me definitively on Marx's side...

Well, I think that that is not quite right the way how to formulate it. It may be that human frailty and disability are so strong that getting rid of the concept of God entirely is not what is going to happen. In the United States we have these hyper-liberal religious denominations or confessions, like the Episcopalians (i.e., the American branch of the Anglicans), the Congregationalists, and the Unitarians, holding that there is at most one God. Most of them are really atheist, but they like to go to church and to have their own community. If religion were like that everywhere, there would be no particular reason to discourage it. So, I am in favour of talking candidly about religion. If we just get used to talking more openly, more candidly, and more factually about religion, getting away from the idea that we are not supposed to talk about these things that would be a breath of fresh air.

REFERENCES

Deacon, T. W. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York: W. W. Norton & Company.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little Brown & Company.

Dennett, D. C. (1996). *Kinds of Minds: Toward An Understanding Of Consciousness*. New York: Basic Books.

Dennett, D. C. (2003). *Freedom Evolves*. New York: Viking Adult.

Giorello, G. (2010). *Senza Dio. Del buon uso dell'ateismo*. Milano: Longanesi.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *359*(1451), 1775-1785.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*(4), 529-539.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49-54.

Report

Truth, Knowledge and Reality

SIFA (Società Italiana Filosofia Analitica) – IX National Congress

University of Padua, Dep. of Philosophy, 23-25 September 2010

Claudio Calosi *
claudio.calosi@uniurb.it

The IX National congress of the Italian Society for Analytic Philosophy was held in Padua, from 23 to 25 September 2010. It was ambitiously titled *Truth, Knowledge and Reality* and has welcomed more than a hundred speakers from all over the world. It has been structured both in plenary and parallel sessions covering areas as diverse as aesthetics, practical philosophy, epistemology, metaphysics, philosophy of science, logic, philosophy of language and philosophy of mind. Needless to say it is impossible even to think of giving a comprehensive account of the conference. In what follows we will then focus on some of the given talks.

We could not but start from one of the most important contributions to the overall congress. In a plenary session Stathis Psillos (University o Athens) and Mauro Dorato (University of Roma 3) have discussed Ontic Structrural Realism[1] and the possibility to add modality in its support.

Psillos in his paper repeats his celebrated critique to OSR. It has been argued, in particular by French and Ladyman, that this critique can be overcome if only we add modality to the picture, i.e we require that structure has an irreducible modal nature. Psillos then rehearses various ways in which modality can be added in support of OSR and finds them all untenable. In particular he argues that the most promising strategy, that of employing so called structural universals, fails on both physical and metaphysical grounds.

Mauro Dorato, in his discussion of Psillos' contribution, has pushed his points even further. He claims that the compatibility claims between physics and metaphysics are indeed all we can ask and that we should stop imposing metaphysical categories, such that of structure, as intended sometimes by

* Urbino University

[1] OSR from now on.

OSR, in order to describe the ontology of natural sciences. Rather we should understand science in its own terms.

In what follows we focus on three different papers given in the parallel sessions. In the first one, *Towards a C-Theory of Time*, Matt Farr (University of Bristol) attempts to construct a theory of time that is a viable alternative to the main celebrated theories, namely A and B theories of time. The main difference between a C-theory of time and its more celebrated rivals is that this theory does away with the notion of directionality of time, understood as directionality of time itself rather than objects and events in time. Farr's main motivation for constructing such a theory comes from physics. Fundamental laws of physics such as laws of classical and relativistic dynamics, laws of the electromagnetic theory, and the Schroedinger's equation governing quantum evolution are all time reversal invariant. This fact, Farr contends, should be adequately reflected in a metaphysical theory about time. The primitive notion of such a C-theory should be Betweness, B (e_1, e_2, e_3), for event e_2 is between events e_1 and e_3. This account leaves open the question of the temporal relation holding between e_1 and e_3 and so it genuinely does away with any directionality of time. A major problem could come for this account when considering measurement in quantum theory. Suppose you have e_1 = a quantum state being $c_1|\uparrow> + c_2|\downarrow>$, e_2 = the quantum state being after a spin measurement $|\uparrow>$ and e_3= the quantum state being after another spin measurement $|\uparrow>$. The Farr's theory of time would be just able to say that e_2 is between e1 and e_3. However, based on our present knowledge of Quantum Mechanics, we would want to be able to say that there is just one possible direction for the quantum system evolution, namely e_1, e_2,e_3.

The next paper we want to focus on is, in our opinion, one of the strongest presented. We are talking about Andrea Borghini (College of the Holy Cross) and Marco Nathan's (Columbia University) Diachronic Identity in Biology and Philosophy. This paper explores four different independent criteria for identifying individuals, i) morphology, ii) function, iii) evolutionary history and iv) development. The authors focus on the fourth criterion that has so far been rather neglected. They present a detailed case study, taken from recent biological studies in the fields of embryonic stem cells[2], in which, they contend, the first three identity criteria, would fail to distinguish individuals. They go on to argue, rather convincingly, that in the ESC case, the

[2] ESC from now on.

development criterion scores better. With these results, learned in close proximity to biological sciences, they rethink classical philosophical problems related to diachronic identity, such as persistence through time and change. Their work is one of the finest example of fruitful interaction between sciences and philosophical reflection at its best.

To conclude we spend two words on another work, namely Claudio Calosi's (University of Florence) *Metaphysics of Persistence and Unrestricted Composition*. In this paper the author sets out to prove rigorously that the endorsement of the rather controversial mereological principle of unrestricted composition, roughly the principle according to which given any two non empty sets of objects there always exist a mereological sum of those objects, dims one particular metaphysics of persistence, namely Three-Dimensionalism[3], wrong. 3D roughly maintains that all material persisting objects are multilocated at temporally unextended spacetime regions. The author constructs a counterexample to such an universal claim using the principle of unrestricted composition. The weakness of this kind of argument is probably that it will appeal to a four-dimensionalist but will not move a three-dimensionalist. She will probably just insists that the argument shows we should not have bought into the unrestricted composition principle in the first place.

This works were chosen just to give a flavor of the entire conference. It covered basically every crucial field in contemporary analytic philosophy and had gathered together leading scholars and young researchers, discussing and confronting different approaches and thesis. And this is, supposedly, philosophy.

[3] 3D from now on.

HUMANA.MENTE - ISSUE 15, JANUARY 2011

Agency: From Embodied Cognition to Free Will

Edited by Duccio Manetti & Silvano Zipoli Caiani

Today embodiment is a critical theme in several branches of the contemporary philosophical debate. The term embodiment refers to the role of an agent's own body in making possible many experiential and cognitive abilities, suggesting the existence of a deep connection between action and perception.

This issue of Humana.Mente will focus on the relationships between our sense of agency and the various models of the mind and of the self. The volume will be organized into two different sections: one concerning the discussion about the theory of agency; the other concerning the theories on free will. The purpose of this project is to collect relevant studies in these fields, opening the door to an interaction between perspectives from various disciplines such as psychology, cognitive sciences, neuroscience and philosophy.

For more information about the journal
visit our website at:
www.humanamente.eu

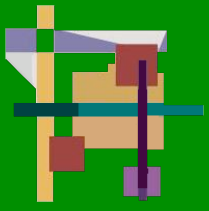
www.humanamente.eu

Humana.Mente – Journal of Philosophical Studies was founded in Florence in 2007. It is a peer-reviewed international journal that publishes 4 issues a year. Each issue focuses on a specific theme, selected from among critical topics in the contemporary philosophical debate, and is edited by a specialist on the subject, usually an emerging researcher with both philosophical and scientific competence.

Humana.Mente wants to be a place for exploring the most recent trends in the international philosophical discussion and wants to give the opportunity to the international community of young researchers to confront each other, and to discuss, control and verify their theories. An analytic perspective is favored, and particular attention is given to the relationship between philosophy and science, without however neglecting the historical aspects of the philosophical topics.

In this issue papers by:

MARK H. BICKHARD (Director of Institute for Interactivist Studies, Lehigh University) - TONY CHEMERO (Franklin and Marshall College) & MICHAEL D. SILBERSTEIN (Elizabethtown College) - TERRY HORGAN (Arizona University) - ROBERTA DE MONTICELLI (San Raffaele University, Milan) - SHAUN GALLAGHER (Florida University) - MAURO MALDONATO (University of Basilicata) - SUSAN POCKETT (University of Auckland) - DAVIDE RIGONI (University of Padua), LUCA SAMMICHELI (University of Bologna), MARCEL BRASS (Gent University) - STEVE TORRANCE (Sussex Uk) & TOM FROESE (University of Sussex) - JING ZHU (Institute of Philosophy of Mind and Cognition, National Yang Ming University)



HUMANA.MENTE
Journal of Philosophical Studies

ISSN: 1972 -1293