Frontiers in Neuroethics

Frontiers in Neuroethics:

Conceptual and Empirical Advancements

Edited by

Andrea Lavazza

Cambridge
Scholars
Publishing

Frontiers in Neuroethics:
Conceptual and Empirical Advancements

Edited by Andrea Lavazza

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-8592-4
ISBN (13): 978-1-4438-8592-8

TABLE OF CONTENTS

INTRODUCTION: THE RELEVANCE OF NEUROETHICS

ANDREA LAVAZZA

Brain sciences are increasingly involved in both scientific research and contemporary culture in general. The great step forward made possible by new techniques of investigation has opened the door to a growing understanding of the functioning of our nervous system. This understanding has allowed us to hypothesize that the material base of the brain, despite being extremely complex, might be the key to explaining human thinking and behaviour and, consequently, also culture and society. "Subdisciplines" such as neuroeconomics or neuroaesthetics are clear evidence of this paradigm shift. All this constitutes a revolution that is still in progress but does not have a central aggregator – and maybe needs one. Neuroethics, I believe, is the best candidate to be such an aggregator.

Although the term “neuroethics” appeared at the end of last century, the development of a new border-discipline, which is still under construction, dates back to the early 2000s.[1] In particular, I am referring to the seminal conference "Neuroethics: Mapping the Field", held in 2002 in San Francisco, as the foundation and start of this field of interdisciplinary reflection (Marcus, 2002). Neuroethics has its hub in the new knowledge on human brain functioning and concerns the moral implications associated with this

[1] The first known occurrence of the word "neuroethics" dates back to 1973, in the work of Anneliese Pontius (1973), who used it in the discussion of the attempts to make children walk before the appearance of the natural tendency to move in a standing position: a practice that can be harmful in the long term. According to Eric Racine (2010), it is a proper use of the term, also in light of the current meaning of "neuroethics", as it was applied to the ethical implications of a form of behaviour "enhancement" and implied the relevance of neuroscientific data to the ethical discussion. Subsequently, Ronald Cranford (1989) introduced the term "neuroethicist" to describe a clinical expert in ethical dilemmas related to neurological issues.

research as well as the application of its results. In little more than ten years, the development has been impetuous, with a growing number of dedicated articles, books, conferences, university departments and scientific societies. The centre of neuroethical research is the English-speaking world, with the quadrangle made up of the United States, Canada, Britain and Australia, but the field is also gaining popularity in continental Europe, where Italy plays an important role.

Neuroethics is a disciplinary field that is still fluid and has elastic boundaries, located at the convergence between some special sciences whose epistemological apparatus is conceptually well-structured, philosophy and ethics (and, within the latter, both metaethics and normative ethics). At present, neuroethics can probably be defined, in Kuhnian terms, a pre-paradigmatic discipline: research produced by the community of scholars has not yet found a specific structure in this field. However, neuroethics is rapidly being organized, both as regards the themes that are part of it and as regards the shared method of investigation.

The interdisciplinary nature of this new discipline makes it a space of intersection between different fields of knowledge – from neuroscience to psychology, from philosophy of mind to molecular genetics and the theory of evolution – which makes neuroethics naturally destined to integrate effectively with another rapidly evolving interdisciplinary field: that of cognitive science. In this sense, neuroethics can be considered to be primarily related to conceptual clarification and a certain epistemological-pragmatic orientation, drawing from the outside the empirical data on which to work. However, it can also have the ambition to experiment directly, while still rejecting, at least in part, the non-evaluative character typical of the "hard" sciences. In fact, on the one hand, there is neuroethical research conducted in the laboratory on moral reasoning and its basis in the brain, resulting from the collaboration between neuroscientists and philosophers; on the other hand, there are ethical reflections on the acquisitions of neuroscience that are related to the more general recovery of normative ethics in the new millennium.

Neuroethics is surely a young discipline, but it is already full of promising developments in various directions. Therefore, it makes sense to propose a collection of contributions that show its current wealth and potential. In particular, the present volume contains papers that focus on some theoretical perspectives and others that provide significant examples of its exper-

imental aspects. Together, they constitute a good introduction to neuroethics as well as a suggestion for further investigations.

The first part of the book includes mainly theoretical studies and presentations of research results. In the first chapter, "Neuroethics: A New Framework. From Bioethics to Anthropology," I seek to outline more accurately the boundaries of neuroethics. Focusing on the interdisciplinary nature of this field of study, I mark its difference from bioethics, highlighting the points of divergence of the two disciplines. In particular, I contend that bioethics seems more focused on what one can do and therefore what one should or should not do. It is an essentially normative task, based on the technical possibilities available. On the contrary, what seems to characterize neuroethics is that it focuses on what we know about the human brain and the way it works. This knowledge inevitably produces some social, political and legal, as well as philosophical and ethical consequences. Neuroethics, I believe, finds its own special area of competence and its specific role in identifying and dealing with such consequences.

In the chapter "The Contribution of Blindsight to the Understanding of Consciousness: Empirical, Conceptual and Normative Implications," Marcello Ienca focuses on blindsight. This term refers to the phenomenon of patients with a lesion in their primary visual cortex who can respond to visual stimuli presented in the blind part of their visual field. Studies on blindsight subjects are a well-established strategy in consciousness research, in particular in the field of cognitive neuroscience. Ienca addresses the questions of how and to what extent the blindsight approach can contribute to the understanding of consciousness. In addition, he outlines the major implications of blindsight research for the neuroscience of ethics as well as for the ethics of neuroscience. He concludes that, although research on blindsight is affected by significant limitations, it contributes to an important degree to the understanding of some aspects of consciousness and poses several non-trivial challenges for neuroscience, philosophy and ethics.

In the chapter "Dualism and Materialism in the Era of Neurotechnology: The Case of DBS," Giulio Mecacci and Pim Haselager focus on deep brain stimulation and its adverse psychological implications based on a relevant number of postoperative situations. Whether these effects have to be attributed to a reactive response to a new situation or whether they are caused by the stimulation itself, or both, remains to be elucidated. Mecacci and Haselager evaluate how various views and conceptual schemes con-

cerning the mind–brain relationship might play a role in the (mal)adaptation following DBS treatment. They hypothesize that the frequently reported maladaptations might be partially caused by a conceptual shift away from dualism and towards a "neurocentric" materialism, promoted by the scientific explanation of the pathology. Finally, they respond to several objections raised against their hypothesis.

To conclude the first part, Marco Fasoli's chapter "Neuroethics of Cognitive Artifacts" deals with the emerging topic of cognitive artifacts. The latter are objects that "shape and transform our cognitive system and cognitive practices" (Heersmink, 2013). Nowadays, this kind of artifact plays an important role in many cognitive tasks that we used to carry out without them in the past. Cognitive artifacts are also changing our behaviours and our cognitive practices. Fasoli shows why, when employing our recent understanding of the human mind and cognition, we should develop a neuroethics of cognitive artifacts. The main goal of this neuroethics of cognitive artifacts will be that of discriminating between right and wrong uses of cognitive artifacts and between right and wrong cognitive designs of complex artifacts, such as technological devices.

The second part of the book gathers papers focusing on experimental research, showing the role of neuroethics in the field of empirical studies. In the chapter "Moral Computation," Alessio Plebe shows how the study of human morality has benefitted from a plurality of perspectives across various scientific disciplines, in addition to traditional philosophical speculations. In light of this, he explores the computational modelling approach. Currently, he maintains, three main strands of research can be identified. The first, with the longest tradition, is based on formal logic. Its main focus is the semantics of sentences containing moral predicates. The second line of research is inspired by generative grammar theory, and is proposed under the name of Universal Moral Grammar. The third stream of research, still in its infancy, aims at reproducing particular brain computations that appear to be crucial in morality. Plebe presents an early example of morally mediated behaviour: a neural model induced to steal food.

In the chapter "Modulating Morality: Can Subthalamic Deep Brain Stimulation Alter Moral Conflictual Decisions?" Manuela Fumagalli and Alberto Priori discuss the possibility that an external manipulation, acting on brain functioning, can alter morality in humans. Their first aim is to define moral conflict according to psychological theories, and to discuss the role of a specific brain structure – that is, the subthalamic nucleus (STN) – in

moral conflict. They also present and discuss the findings of the only study that, to their knowledge, has tested the ability of the electrical stimulation of the subthalamic nucleus to alter moral judgment in patients with Parkinson's disease. These findings raise relevant ethical questions and pave the way to further research.

In the chapter "The Ethical Ghosts in the Machine: Testing the Relationship between Consciousness and Responsibility in the Special Case of REM Sleep Behaviour Disorder," Matteo Cerri focuses on the REM Behaviour Disorder (RBD). This disease presents some very interesting features that can be used to test the interpretation of the relationship between consciousness and responsibility. In fact, RBD is a parasomnia of REM sleep: a state where the presence of conscious activity can be hypothesized. According to the Integration Information Theory (IIT), during REM sleep, the cerebral cortex presents the necessary activity to support consciousness. If we can consider ourselves conscious during REM sleep, how should we judge the actions performed in such a state? Cerri argues that a closer look at the relationship between consciousness and responsibility during an RBD episode may enrich, and potentially change, our idea of responsibility.

Finally, in the chapter "Are Near-Death-Experience Memories Real? Ethical Implications of a Neuropsychological Study," Johann R. Kleinbub and Silvia Zidarich focus on the distinction between real and imagined memories. The authors try to push the boundaries of this research field by applying its methods to the memories of near-death experiences (NDE), a phenomenon of still unknown nature, characterized by a strong "phenomenological certainty," typical of the perception of daily life events. NDEs have been described by individuals as "real" and often as the most intense, vivid, important, and founding experience of their lives. Kleinbub and Zidarich adopt an integrated approach involving a hypnotic procedure to improve the recall process together with EEG recordings in order to investigate the characteristics of memories and their neural markers comparing memories of NDE, real and imagined events. Their results are discussed in the light of the ethical and epistemological implications of researching the "boundaries of life."

Taken together, the eight chapters offer a good presentation of neuroethics as a lively and dynamic field of study, hopefully providing the reader with an exhaustive introduction to ongoing debates about mind, brain, con-

sciousness, and the ethical issues related to the findings in neuroscience, both molecular and cognitive.

References

Cranford, R., 1989, "The Neurologist as Ethics Consultant and As a Member of Institutional Ethics Committee," *Neurologic Clinics*, 7, pp. 697-713.

Heersmink, R., 2013, "A Taxonomy of Cognitive Artifacts: Function, Information, and Categories," *Review of Philosophy and Psychology*, 4, pp. 465-481.

Marcus, S.J. (ed.), 2002, *Neuroethics. Mapping the Field*, Dana Press, New York.

Pontius, A., 1973, "Neuro-ethics of 'Walking' of the Newborn," *Perceptual and Motor Skills*, 37, pp. 235-245

Racine, E., 2010, *Pragmatic Neuroethics: Improving Treatment and Understanding of the Mind-Brain*, The MIT Press, Cambridge (MA).

PART I –

THEORETICAL PERSPECTIVES

NEUROETHICS: A NEW FRAMEWORK— FROM BIOETHICS TO ANTHROPOLOGY

ANDREA LAVAZZA

1. Studying "who we are" also thanks to neuroscience

If it is true that neuroethics is still a fluid discipline, located at the crossroads between special sciences with a conceptually well-structured epistemological apparatus, it is also true that it is somewhat still in search of a precise object of study. In this sense, a useful and often referred-to distinction is the one proposed by Adina Roskies (2002) between the ethics of neuroscience (clinical neuroethics) and the neuroscience of ethics (the area of study dealing with the way in which ethics is rooted in the brain). This dichotomy grasps different and complementary aspects. In fact, "the ethics of neuroscience" regards the reflection on the controversial applications of neuroscience itself (all those aspects requiring a moral judgement), while "the neuroscience of ethics" focuses on metaethics: that is, a study of moral reasoning based on its cerebral roots. Such a view can be also found in Farah (2005), who proposes a division between the "practical" and the "philosophical" aspects of the discipline.

More recently, Eric Racine (2010: 5) has proposed a quadripartition to better outline the potential areas of competence for neuroethics:

(1) Research Neuroethics: the ethical challenges we face through neuroscience (for instance, the theme of the use of embryonic stem cells, or the use of animals in experiments; or even the accidental discovery of clinical abnormalities in human subjects examined for scientific purposes);[1]

[1] The examples are mine, not Racine's.

(2) Clinical Neuroethics: the challenges related to treatments for neurological or psychiatric patients (the safety and availability of treatments; side effects,[2] etc.);
(3) Public and Cultural Neuroethics: the sphere regarding the new understanding of pathological conditions and their social dimension (what does it mean to suffer from a psychiatric disease, the new understanding of depression as an organic pathology rather than an existential condition, etc.);
(4) Theoretical and Reflective Neuroethics: the theoretical and epistemological bases of the discipline itself, as well as its general implications for ethics and some moral concepts about the way in which we judge both practically and theoretically (is the mind-brain equivalency grounded?; why and when are we led to follow a utilitarian criterion?; etc.).

Racine's is a very wide description, rather blurry in its boundaries, including problems new and old – some of which are dealt with by bioethics and moral philosophy, whose subsumption under a common label might be regarded as useless or misleading. On the other hand, it is hard to deny that cognitive neuroscience, researching the brain correlates of so-called superior human functions, aspires to occupy the space of psychology and philosophy – in fact, this process is already in place. Colin Blakemore writes:

> As our understanding of brain function advances, it is surely reasonable to ask how that knowledge illuminates issues that have previously been expressed, formulated, or explained in other terms. Epistemology – the theory of knowledge, legal principles, social and political concepts of rights and responsibility, religious beliefs and behaviour, the philosophical underpinnings of science: all of these are products of our brains, and they deserve consideration within the framework of our knowledge of the brain (Blakemore, 2006: V).

Surely this statement is still too general. To say that "all of these are products of our brains" is to say something true but not particularly informative, given that a functioning brain is the necessary condition for any cognitive activity. However, the single individual's brain might not be enough to explain social and political concepts of right and responsibility, religious belief and behaviour, and the philosophical underpinnings of science. Here, what comes into play is culture, understood as externalised

[2] For instance, treatments for Parkinson's disease may cause compulsive gambling.

knowledge, the interaction between different brains, and institutions as rules to which to conform. However, the growing knowledge of the structure and functioning of the central nervous system and the consequent increasingly close "coupling" between its activities and mental functions – thanks to microscopy, techniques of brain imaging and magnetic stimulation, and genetics – has produced a capacity to intervene on the brain that goes beyond the treatment of disorders once considered exclusively organic (neurological), distinguished from generally psychological and psychiatric ones.[3] Also, now we can also act on a "healthy" or "normal" brain by changing the subject's behaviour – which leads to entirely unprecedented scenarios. Hence neuroethics in its first sense.

> Neuroethics – the examination of what is right and wrong, good and bad about the treatment of, perfection of, or unwelcome invasion of, and worrisome manipulation of the human brain (...). What ethical rules or legal regulations should there be for treatment to change criminal behavior? Should we develop a drug to improve memory or to repress painful remembrances? Or to help a prosecutor elicit a professedly forgotten detail? Is it fair to implant a chip in the brain to enhance memory before an academic examination? (…) Is the imaging of suspected terrorists' brains to detect lying a form of torture, or at least a way of forcing people to incriminate themselves? (Safire, 2002: 5-8).

In light of this, the first symposium specifically dedicated to neuroethics was organized, setting the stage for reflection on the ethical, legal and social implications of neuroscience (summarized in the acronym *Elsi*). This is the "classical" setting of neuroethics as it appeared in its first years. The implications and consequences of neuroscience – which, from the outset of the discipline, have attracted the attention of scholars (and the public), from issues of brain privacy to upgrade interventions – have been almost exhaustively addressed (Marcus, 2002; Gazzaniga, 2005; Illes, 2006; Glannon, 2007a; 2007b; 2011; Levy, 2007; Giordano and Gordijn, 2010; Racine, 2010; Farah, 2010; Illes and Sahakian, 2011; Chatterjee and Farah, 2012; Clausen and Levy, 2014; Boella, 2008; Lavazza and Sartori, 2011; Sironi and Di Francesco, 2011; Reichlin, 2012; Corbellini and Sirgiovanni, 2013; Evers, 2012). Some of the questions that neuroethics has initially set are certainly new for the techniques and the knowledge about the brain they involved, but they still may be answered with the conceptual and

[3] Some of these topics have already been addressed in Lavazza (2011).

normative tools that moral philosophy and bioethics in particular have developed in recent decades.

What I want to propose here is a different focus, one that characterizes neuroethics with greater specificity and makes it a field of interdisciplinary research with its own identity and its own object. As for the methodology, neuroethics should adopt that of the disciplines that make it up and whose variety is essential to the project of neuroethics that I propose. In some previous papers (both in English and Italian) I have already introduced some of the ideas I will develop here (Lavazza, 2011; Lavazza and De Caro, 2013; D'Alessio *et. al*., 2014).

More recently, in the major collection of works devoted to neuroethics so far (Clausen and Levy, 2014), the discipline has been defined "as a multidisciplinary and inter-disciplinary endeavor" that "examines the implications of the neurosciences on human beings in general and on their self-understanding and their social interactions in particular." Such ideas, however, were not taken as the guiding motive of the work, which rather includes a "range of approaches adopted in neuroethics, including historical, anthropological, ethical, philosophical, theological, sociological and legal approaches," somewhat going back to (and even widening) the general setting of the first dedicated handbooks, and also including specific aspects of neuroscience that seem less central, if not peripheral, to neuroethics.

I am clarifying this because, based on the approach I am proposing here, neuroethics has as its own object (and should focus on) *what we learn about ourselves and the way we innerly "work"* mainly, but not exclusively, thanks to neuroscience. In other words, it is the *strong* naturalization of the human being that calls for a *discipline* resorting to different existing forms of knowledge, trying to integrate them. Thus, its object of study is not *what we can do*, but *what we know or reliably believe we know*. In fact, unlike bioethics (which may aspire to prescribe or prohibit things), when it comes to understanding "how we are made" we cannot put the "genie" back into the "bottle": once available, knowledge of the functioning of the mind has inevitable philosophical effects related to human self-understanding, with social, political, legal and economic consequences. This also explains why a reflection on neuroethics implies *ipso facto* a reflection on the philosophical bases and consequences of neuroscience. All of this is strictly linked to the complementary side in which evolutionary biology, psychology and cognitive neuroscience converge: in it we question the "mentalist" representation we have of ourselves, reversing the intuitive conception of subjectivity as unitary and transparently accessible to

introspection, as well as the explanatory and causal value of intentional psychology.

There are still many things left unexplained on this empirical front, with room for philosophical inquiry as such. We might recall here what Jürgen Habermas (2013) proposed for other fields: for him, philosophy can act as a process of self-clarification of the Self. If scientific disciplines focus on a given objective sphere, philosophy considers the outcomes of scientific learning and the effects that knowledge of a part of the world might have *for us*. Philosophy works on a ground where the new understanding of the self and the world – says Habermas – go hand in hand. The objectivity of the scientific work of philosophy thus lies in the generalizations it produces. The Self of philosophical self-understanding is an abstraction that can be expressed as "for us, for human existence in general." That is, I believe, a form of anthropology.

That's why "neuroethics" (broadly understood, so as to include a philosophy of neuroscience) might be an appropriate name for a new perspective of reflection and research that is not solely related to neuroscience, but has a special focus on it, including other sciences studying the human being in its natural dimension – from genetics to psychology, from social and human sciences to philosophy. As I mentioned, neuroethics concentrates on the methodological aspects, the conceptual clarification and unification of particular and specific perspectives, analysing the social and pragmatic consequences (transitory as they may be) of new forms of knowledge, as well as the meta-ethical aspects related to them. For this reason, neuroethics is proposed as overarching the sub-disciplines characterized by the prefix "neuro," and also as a (meta) discipline bridging areas of hyper-specialization.

2. An example of new neuroethics: is there such a thing as the "self"?

Let's now consider an example related to the fruitful interaction between empirical research and philosophical-conceptual reflection (in fact, as I have mentioned, theoretical psychology and philosophy can "guide" neuroscience at a conceptual level, whereas empirical research helps philosophy and psychology progress thanks to the information deriving from its experimental data). The theory I wish to consider and critically assess is the deflationist theory of the self proposed by Thomas Metzinger (2003; 2009). I believe it is particularly apt for representing what, to me, is the

central sphere of competence of neuroethics, according to the partition I am suggesting here.

In fact, Metzinger, starting from observations about the functioning of the nervous system and from phenomenological descriptions of the self, comes to formulate a hypothesis that is both factual (as regards the observable aspects of the nervous system) and ontological (as regards the existence of an entity such as the self). From his conclusion, i.e. the absence of a real self, derive important practical and normative consequences: in fact, this notion affects our conventions and behaviours, as we cannot pretend to ignore what we have discovered, and have to adjust our previous knowledge and consequent conduct to it.

Metzinger's (2003) main ontological thesis, in the framework of what he calls SMTS (the *Self-Model Theory of Subjectivity*), is that *in the world there are no such things as selves*. In his view, what is generally called a "self" is not a substance, an unchangeable essence (i.e. an individual in a metaphysical sense) but a special type of representative content: it is the content of a model of the self that cannot be recognized as a model of the system using it. So, the dynamic content of the phenomenal self-model is the content of the conscious self: our bodily sensations, our present emotional situation and the cognitive contents we experience phenomenically are, for Metzinger, the constitutive elements of the phenomenal self-model. It is an integrated process and not something substantial or otherwise identifiable. So, we can conclude that we are our phenomenal self-model.

This means that we, as conscious beings, do not have and are not a self, even though, for Metzinger, we can fall into inevitable confusion because of the way we are made. Therefore, a true picture of what exists should portray biological organisms that have models of the conscious self, provided that these models are not selves in the full (ontological) sense, but rather complex brain states. If one acts on the basis of a transparent model of the self, then the body has a phenomenal self, which, however, is only an internal and dynamic representation of the organism.

In this sense, the phenomenal experiences of *substantiality* (being an independent entity), *essence* (defined by the presence of an inner core that cannot be modified, and by a set of invariant intrinsic properties), or *individuality* (being a unique and indivisible entity) are only special forms of consciousness. And this also applies to the experience of *perspectivalness*, i.e. the first-person point of view given by a model of the self centred on a

single, coherent and continuing phenomenal subject with its own, not only material, positionality. According to Metzinger, what causes a representational property to become a phenomenal property of the self is the requirement of *transparency*. The phenomenal self, i.e. the illusion of the substantial self, emerges from a naive-realist self-misunderstanding. And such misunderstanding is caused precisely by the transparency, which is a phenomenological and non-epistemological concept deriving from a lack of knowledge.

In other words, as stated by Metzinger himself, transparency is a special form of darkness. Phenomenal transparency is the fact that the early stages of information processing in the brain are not available or accessible in any way during introspection. The instruments of representation cannot be represented as such, and therefore the system that performs the experience is trapped in a naive realism that cannot be transcended. At the epistemological level, one can speak of "epistemic closure," i.e. a structural deficit in the ability to obtain knowledge about oneself. Transparency, therefore, is given if the previous processing steps are unavailable to attentional processing. Transparency is the product of a structural-functional property of the neuronal processing of information that takes place in our brain, which makes the previous processing steps unavailable to (conscious) attention.

The evolutionary explanation is that the brain has produced "transparency" to avoid the system's regress *ad infinitum* in self-representations. In fact, if a system kept accumulating levels of self-representation, it would end up exhausting the computational resources available and would be bound to paralysis. The self is then reified as a result of a natural process – one that has always taken place and still takes place exclusively in biological terms in the brain, which interacts with the world – that provides the system with information on itself without the need for an internal cycle of self-modelizations built one on the other. However, the cerebral base of the process does not mean that we cannot study it at the neurobiological, computational and phenomenal (that is, philosophical) levels.

The relevant point here is that for Metzinger only brain facts exist, and it is from his specific interpretation of brain facts that his eliminative theory of the self derives. For instance, he derives from this interpretation the need to identify a common functional core to global states like wakefulness and dream. The counterintuitive result is that conscious wakefulness would be nothing but a dream-like state, with the only constraint posed by specific sensory inputs that reach the brain from the outside world. Based on the neuroscientific research by Llinás e Paré (1991), Metzinger states that

what we experience subjectively when we experience our world as consistent is the high degree of internal correlation existing between a subset of physical events that occur in our brain. The conscious model of reality, then, is essentially an inner construct, which is only disrupted by external events that force it to continuously reorganize itself in new stable states.

There are different degrees of transparency. Total transparency is that of animals endowed with attention but not awareness, whereas humans are characterized by certain opacity. This opacity is manifested in the possible partial breach of the auto-epistemic closure. The fact that a part of our self-model is opaque allows us – in Metzinger's opinion – to conceive the possibility of an appearance / reality distinction, not only in the case of our perceptual states, but also as to the content of self-consciousness. It allows for self-distancing, which enables us to critically evaluate the content of any phenomenal self-model; also – through the opaque simulation – it allows us to conceive the epistemological possibility that every phenomenic representation really is a simulation, if objectively considered. Finally, it allows us to conceive the possibility that every phenomenic self-representation is truly a self-simulation.

Here's an interesting point from a specifically neuroscientific perspective: can Metzinger's objectivist and naturalistic turn, for which phenomenal self-models are wrong and selves are ontologically inexistent, somewhat retroact on the process giving life to the phenomenal self-model itself? Metzinger rightly claims that the basic architecture of our phenomenal space changes neither when we think in terms of our opacity, nor when we resort to objective tools for an external analysis of our brain (for example, think of the well-known experiments conducted by the Metzinger himself on out-of-body-experiences and illusions concerning the whole body).

However, one may wonder if the empirical knowledge of the fact that our self is an illusion (and that particular type of illusion) might interfere with the realist "intuition" that comes from transparency. The fact that a stick partially immersed in water always seems bent does not prevent educated people and scientists in particular to almost automatically ignore the optical illusion consubstantial with the human visual system. They know how to "make up" almost automatically at the cognitive level the illusion of the bent stick and, despite the visual impression, it does not affect them in any way.

So, once the idea of the phenomenal self-model and transparency is common, could it end up disrupting or contrasting the rise of the illusion of the

self? The consequences would be clearly relevant to a number of areas of our social, relational, political and legal life. In fact, denying the existence of the self is denying what we otherwise call "I," and this ultimately amounts to denying the existence of me writing, or you reading these lines. But if you and I don't exist, then humans do not exist either, given that humans coincide with people like you and me. And if humans do not exist, many of our interactions risk ceasing to exist or at least being revisited.

However, in my proposal, neuroethical research does not merely discuss, analyse and normatively evaluate the implications of neuroscientific knowledge, theories and hypotheses, but it actively participates in their elaboration and testing. In this sense, Metzinger's theory can be set against the one proposed by experimental psychologist Stanley B. Klein (2014), who deals with several forms of cognitive diseases. According to him, the self is real, causally effective and made up of multiple aspects with different roles in experience and behaviour (Klein, 2014: VIII). What is interesting here is especially Klein's defence of the existence of a self with "classical" features both on a functional and on an epistemological level. Klein substantially distinguishes between an epistemological self (instantiated by the brain) and an ontological self (first-person subjectivity): those two entities interact and cannot be unified in a common lower-level substratum.

In particular, the epistemological self is a neuro-cognitive system of the person in a psycho-physical sense, and it can be the object of scientific research. According to Klein, what identifies the epistemological self is its contribution in terms on knowledge of the world to the benefit of the ontological self. Because many of its processes are unconscious, it emerges when its contents are relevant to the self-consciousness of the ontological self. What characterizes the epistemological self is the sense of personal belonging of mental states, the fact that experience is felt as one's own and linked to the self – therefore, contrary to what was posited by Hume, it does not amount to the sum of single unrelated percepts and experiences.

Indeed, some clinical cases show that some people may be well aware of being the authors of a given experiential content; however, this knowledge does not create *ipso facto* the feeling that such content belongs to them: rather, such content is accompanied by a feeling of detachment. For Klein, personal experience is a sort of "mental glue" creating a sense of uniqueness: "What makes a state distinctly and uniquely 'mine' is that I intuitively sense – without need for intuition, inference, or reflection – the connection between content in awareness and the *feeling* (not the *knowledge*) that this content is uniquely and infallibly my own" (Klein, 2014: 16).

The *epistemological self*, which has long been the object of clinical observations and systematic experimental studies, seems to be made up of a number of different, functionally isolatable neuro-cognitive systems. In particular: 1) episodic memory of one's life events; 2) semantic summaries of one's personality traits; 3) semantic knowledge of facts about one's life; 4) an experience of continuity through time (the "I" experienced now is connected to the "I" experienced at previous and later points in one's life); 5) the physical self: the ability to represent and recognize one's body; 6) the emotional self: the ability to experience and produce emotional states that provide value, affective valence, and evaluative direction to our actions and reasoning (Klein, 2014: 21-22).

In normal individuals, these elements contributing to self-knowledge cooperate to create the feeling of a subjective unity. For Klein, however, if taken individually, none of them is either necessary or sufficient to experience the self as a subjective individual point of view. The first three listed above, in fact, can be damaged without there being a corresponding loss in the ability to experience the self. What matters, as testified by several psychiatric and neurological cases, is that people devoid of access to the bases of knowledge making up the (empirical) self, still retain the sense of personal identity and subjective unity. Contrary to other reconstructions of empirical findings, it seems that the semantic knowledge of the traits of one's personality is preserved even in patients severely amnesic with regards to episodes of their lives. For instance, Klein talks of a patient who, as a result of anoxia, had lost knowledge (memory) of his daughter's personality traits, but not of his own. Thus, the dissociation in subsystems of our cognitive architecture does not mean that the self is divided nor does it mean that it turns out to be a substantially false construction. Rather, it seems that some subsystems *really* constitute what we usually call "I" or "self" in the traditional sense and that it is not only an illusion that occurs phenomenologically.

Patients in whom the sense of the self is profoundly damaged due to organic causes manifest an interesting phenomenon, unrelated to introspective reports (which Klein considers not sufficiently reliable). Even those who lose the sense of their epistemological self (as appears from behavioural observation) can still experience and express the confusion they feel in their mind, asking what is going on, feeling baffled and afraid because of it. Also, cases of true dissociation between mental contents that are known to be in one's head and the feeling of their belonging to one's self indicate, according to Klein, that there is a functional autonomy between the two selves and that their functional union is not a given – in fact, some

patients manage to appreciate their "ownership" of those mental states through a process of inference. Empirical data thus shed light on a cerebral core of the self that is much stronger and persistent than other kinds of research have been willing to admit.

If all this is true, the consequences will be very different from those deriving from Metzinger's theory. If the self exists in some form, then the practices presupposing it would not be affected by new knowledge on cerebral functioning. However, it must be said that Klein's description of the self is very different from the "substantialist" and psychological one provided by common sense, for which the self is the person's "core" staying basically the same over time despite some gradual (but sometimes even sudden) changes. In any case, both Metzinger's and Klein's theories, starting from a series of empirical data that can be differently interpreted, have strong philosophical and pragmatic consequences.

As this example shows, neuroethics can work both on the empirical and on the philosophical/normative front of personal identity, with a direct and inescapable reference to cerebral bases. In this way, its task would be located at the beginning and at the end of the phase of experimental research: on the one hand, conceptually guiding neuroscientific experiments, on the other hand helping interpret them towards the generalizations "for us" I have mentioned above.

3. The "pragmatic" side of new neuroethics

But if neuroethics focuses on the issue of knowledge of the self and us, in the anthropological perspective I am trying to outline, one of its privileged areas will also be to specify the new factual frames within which to make decisions affecting human conduct and treatment. These factual frames descend from new knowledge about our "functioning" made available by cognitive neuroscience and related disciplines.

Consider a real case, one that in retrospect seems like an ethical experiment conceived on purpose. A Swedish man, named Oscar, was for many years a staunch activist of the vegan movement, which intends to respect all forms of animal life banning from human consumption all products that do not come from the plant kingdom (including milk). ("The reluctant vegan: the case of an older man in a Swedish care home," in Banks and Nøhr**,** 2012). At age 75, Oscar had the misfortune of getting Alzheimer's disease and being admitted to a specialized facility for severe dementia, due to which patients gradually lose memory and other cognitive func-

tions. In line with his previous habits and at the behest of his wife, Oscar is served strictly vegan meals. One day, however, Oscar is mistakenly given a portion of meatballs in tomato sauce which he likes very much, something that also makes him realize for the first time that the dishes he is given are different from those served to other patients. From that moment Oscar refuses to eat vegetables. This is a problem for the staff, divided between the desire to please the patient and the demands of his wife, according to whom the husband really "wanted" to eat vegan and was only distorted by his disease.

This example helps evaluate a series of ethical scenarios related to autonomy and the treatment of subjects strongly impaired in their intellectual abilities. In Oscar's case, the ethics committee finally decided to follow his non-vegan "choice." Beyond some of the bizarre theses that have been proposed – such as that this case proves that veganism is against nature and that Oscar "really" likes meat – there are many reflections to be made. Certainly, many vegans like meat, just as a hermit monk would probably like sex, but obviously this is a truism that tells us nothing about the choices made for reasons other than sensual pleasure and the capabilities required to fulfil them.

To help us solve cases like Oscar's, neuroethics can do something unprecedented, uniting clinical data and ethics (the respect of the patient's autonomy and dignity). Let's begin with the concept of autonomy as it is usually understood in moral philosophy. The concept of autonomy "is generally understood to refer to the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not as the product of manipulative or distorting external forces" (Christman, 2009). Those who are autonomous can decide for themselves without interference from others or personal limitations; they can act according to a project of their own, designed without constrictions. Autonomy also concerns the freedom to decide what to believe in and the ability to weigh up the pros and cons of a given course of action.

Another fundamental element regards the awareness of the rules one establishes and follows; a central role in this is played by rational reflection, i.e. the ability to assess existing norms of conduct and traditions and to choose, with the necessary balance between rational and emotional aspects, which ones to follow and which to ignore. Another way of putting this is to say that being autonomous means having the critical ability to pay attention to the outcome of one's deliberations and being able to be driven by one's purposes. Being autonomous means being oneself, follow-

ing one's own considerations, desires, conditions, and characteristics. It has to do with being what can be considered one's "true self" – which, as such, becomes an indispensable value. Autonomy, as defined thus far, is a philosophical concept that is essentially based on a model of folk psychology sufficient to ensure the above-mentioned conditions of competence – a "naive" psychology, integrated by cognitive psychology and the acquisitions of neuroscience. For this very reason, the case of Alzheimer's, in relation to autonomy, seems to constitute a good test of neuroethics in the sense described.

The main characteristic of Alzheimer's disease is the development of multiple cognitive deficits, which include in particular the impairment of memory and at least one of the following cognitive disorders: aphasia, apraxia, agnosia, or a change in executive functioning. In order to have an established diagnosis, cognitive deficits must be severe enough to cause impairment in the patient's social or occupational functioning. For our purposes, it seems that the fundamental question is the following: in our attempt to respect a patient suffering from dementia, should we stick to the way they acted before contracting the disease or the way they are now? (Jaworska, 2006). There are two established antithetical positions with regards to this: those proposed by Dresser (1986) and Dworkin (1993). According to Dresser, we must follow the patients' perspective the way they manifest it at the time being, given that their previous values are no longer applicable. On the contrary, Dworkin believes that only desires expressed in full autonomy have relevance. Finally, a third option is proposed by Jaworska (2006), who draws on neurobiological findings to contest some of Dworkin's assumptions about the autonomy of people suffering from Alzheimer's.

Dworkin's idea is that to develop what he calls "critical interests" (that is, general purposes even outside of self-realization, such as the well-being of one's children, the relevance of one's work to the community and so forth) – as opposed to "experiential interests" (related to momentary personal satisfaction) – one has to be able to see one's life in a unitary way, linking the past to the present and the future, which is a skill that Alzheimer's patients don't have. Devoid of this condition of autonomy, they cannot generate new critical interests. However, as Jaworska objects, it is legitimate to refer to a different conception of autonomy, one for which dementia sufferers continue to possess the mental functions on the basis of which they are able to generate new critical interests.

Such autonomy amounts to the ability to make assessments, judge values as positive and right and adhere to them, being able to give some reason for their preferability. Think of a dementia patient that, when questioned, cannot correctly indicate the present date, where she lives or how old she is, but volunteers for experiments and medical tests on the grounds that she would help her community. The latter skill does not seem to require an overall awareness of one's life, although it involves some coherence between basic attitudes.

The findings of biopathological analyses support such a distinction. The early stages of evolution of Alzheimer's selectively affect the hippocampus, a crucial region for memory processes: the patient gradually loses the ability to retain the memory of recent events, while the memory traces of the distant past remain more stable. If the anterograde amnesia breaks the narrative continuity of existence, other brain regions are initially spared and do not prevent the emotional attachment to specific courses of action. A confirmation of this comes from known cases, described for example by Damasio (1994), in which patients with injuries to the ventromedial prefrontal cortex, able to pass psychological tests of memory, language, attention and intelligence, are however unable to choose between courses of action despite understanding their features and consequences.

In that case, there is a deficit of decision-making due to reduced emotional involvement: without the activity of that specific brain area, the individual does not develop the typical coloring of pleasure or repulsion that characterizes any choice. In essence, what's missing is the ability to express an assessment or to assign a value. This ability lasts rather long in Alzheimer's patients, because of the different regions affected by the two diseases. In the light of this, Jaworska is able to object to Dworkin's idea that only critical interests prior to the disease are valid: in fact, a thorough examination of the neurophysiological conditions shows that people suffering from dementia still are able to commit to values and critical interests, be they new or different.

According to Dworkin, Alzheimer's patients have no autonomy, understood as the ability to express one's character in one's life, as they no longer have the brain resources to translate values and convictions into practical conducts. Since respect for people's autonomy is a moral imperative, Dworkin concludes that we must stick to commitments that the patients had prior to their illness. For Jaworska, instead, it seems possible to conceive of autonomy differently, by focusing on general purposes, no matter if the subject can implement them through a chain of means fit to

his or her ends. This latter skill is what Alzheimer's patients lose, but this does not amount to a complete loss of autonomy. Even in illness, a subject can still have certain independence, despite his or her behaviours being often inadequate to this aim. Others can help him or her choose and act according to his or her wishes, without this entailing a lesser degree of autonomy. The ability to adhere to a value as a basis for one's wishes and as an assessment of what is right and appropriate can be thus regarded as a form of autonomy, one that Alzheimer's patients still have.

Here it is not necessary to further specify the attribution of autonomy in a case like that of the above-mentioned patient willing to participate in clinical trials. Rather, let's go back to Oscar. The dilemma "does he really want meat?" becomes a neuroethical question with an "empirical" side implying the assessment of Oscar's ability to express a judgment, to evaluate something as positive and right, being able to motivate his choices. In the specific situation it would have been helpful to verify if Oscar could articulate the reason for his sudden preference for meat over vegetables. Maybe even a simple exam to evaluate the status of atrophy of his brain areas would have helped the ethical committee.

However, new knowledge on Alzheimer's and its effects at the cognitive level further complicate Oscar's case. Some studies have focused on the *perception* that close relatives and friends have of the patients' identity disruption (Strohminger and Nichols, 2014; 2015). According to Strohminger and Nichols, it's not mainly memory loss that makes someone seem like a different person, nor are personality change, loss of higher-level cognition, depression or the ability to function in daily activities. Moral traits — more than any other mental faculty — are instead considered the most essential part, the core component of our identity. And injury to the moral faculty plays the primary role in identity discontinuity. The findings suggest "that folk notions of personal identity are largely informed by the mental faculties affecting social relationships, with a particularly keen focus on moral traits" (Strohminger and Nichols, 2014). Such notions "mark a departure from theories that ground personal identity in memory, distinctiveness, dispositional emotion, or global mental function" (Strohminger and Nichols, 2015).
Of course, friends and relatives are no experts in evaluating the continuity of personal identity, but they are the only ones who've known the patients for long enough to make such an assessment to begin with. Their role – with the help of experts and following a strict protocol – is therefore fundamental.

If it is possible to attribute greater identity continuity to Alzheimer's patients as a consequence of the persistence of moral traits, contrary to Dworkin's position, other studies seem to go in an opposite direction as to Oscar's case. In fact, we know (Caramazza and Shelton, 1998; Capitani *et al*., 2003) that Alzeheimer's patients undergo high-level and category-specific impairments in conceptual domains. In particular, there is a category-specific impairment in the conceptual domain of living things. It has been noted (Zaitchik and Salomon, 2009) that in tasks requiring tracking an object's identity in the face of irrelevant but salient transformations, AD patients fail to recognize the species as the same. This fact indicates that they are impaired in their theoretical understanding of living things while they can still track a person's false beliefs in the face of changes affecting the truth of those beliefs. In other words, Alzheimer's patients can retain basic theory of mind while losing the ability of recognizing animals.

This could mean that because of the disease Oscar – whose care for animals was particularly strong, as his veganism shows – lost the specific cognitive resources necessary to that care. If this is the case, it seems difficult to grant Oscar autonomy when it comes to his new food choice. His wife's paternalistic demand, aimed to preserve her husband's previous critical interests, might be in the right as Oscar has no longer the ability to develop new ones. Selective and hard to identify deficits caused by AD seem to lead in this direction.

Finally, if the patient had only been able to say "because I like it," there would have been reason to regard him as a "wanton." According to Harry Frankfurt (1971), a wanton is a creature that has desires of a first order, which are satisfied in an irresponsible way. Many follow their passions without thinking, but also have second-order wishes: that is, they are able to desire in a reflective way what they naturally like. Maybe Oscar had lost that ability: he liked meat, but no longer had the ability to decide not to eat it for the sake of animals. Giving him meat, then, wasn't probably the best way to respect him as a person who, before being ill, expressed specific choices with regard to his moral conduct. Of course there is much to discuss about this. The contribution of neuroethics, which can help us make decisions based on facts and convincing reasons, is to clarify the overall framework as exhaustively as possible, "updating" classical moral reasoning in the light of new knowledge.