Philosophy of Mind

Philosophy of Mind:

Contemporary Perspectives

Edited by

Manuel Curado and Steven S. Gouveia

Cambridge
Scholars
Publishing

Philosophy of Mind: Contemporary Perspectives

Edited by Manuel Curado and Steven S. Gouveia

This book first published 2017

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): 1-5275-0000-4
ISBN (13): 978-1-5275-0000-6

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

INTRODUCTION

The recent surge of interest in Philosophy of Mind has seen the emergence of a multidisciplinary field as a legitimate academic discipline. The acceptance of scientific knowledge as instrumental in solving philosophical problems is one of the key features of this revitalized area of study. This book is a proof of this idea – here you will find several overlaps between philosophy and other areas of knowledge, such as cognitive neuroscience, artificial intelligence or psychology, which can lead to some encouraging advance in several problems of the area.

This collective book seeks to piece articles addressing contemporary Philosophy of Mind's broadly considered issues and problems together. The project started to be conceived within the context of the conferences presented at an international symposium on Philosophy of Mind at the University of Minho (Portugal), and made its way by opening up to the international community through a call for papers, by which some excellent works were received and considered by the organization. It assembles graduate students, junior researchers and senior scholars with an outstanding reputation.

The book will certainly have enough material for researchers in the field (and related areas, such as cognitive science and artificial intelligence) but may also be useful for students of any course of study or degree. It can be used as a guide to some courses at various levels, from BA to MAs and PhD courses of various fields, as well.

The volume is structured as follows: part I will be focusing on one of the most important concepts of this area, the "self"; part II, on sensory perception – particularly odours, colours and vision; part III raises some questions about the future, the ethics and the costs of artificial intelligence; part IV aims to demonstrate how philosophy of mind can benefit from cognitive neuroscience; and part V will consider several historical influences and concepts of the discipline.

***

Part I opens with a paper by Sofia Miguens (University of Porto) with the humean question "Where Am I, or what?" and several difficulties that arise when one tries to answer it. The main concern has to do with

numerous examples that show a lack of authority in self-identification. In this case, the author will consider if there is "any way one is simply unable to go wrong when encountering oneself" and will try to answer it focusing on two points of view: the language-based account of the subjective as first-person authority by Donald Davidson, and the phenomenological account of immunity to error through misidentification in proprioception by Shaun Gallagher. The main goal is to try to answer Hume's question without "looking inside oneself for a 'self'".

The next article, by Thomas Metzinger (Johannes Gutenberg-Universität Mainz), will be centred around empirical perspectives of the self-model theory of subjectivity, presenting several examples of it. The author uses a series of empirical examples from various disciplines – a perfect illustration of the multidisciplinarity of contemporary Philosophy of Mind – to demonstrate the explanatory power and main ideas of the Self-Model Theory.

Finally, to close Part I, Luca Forgione (University of Basilicata) will concentrate on the problem of the knowledge of one's mental states revolving around the involvement of the self-conscious subjective dimension. The author will conclude that the basic capacity for self-consciousness relies on the possibility to produce I-thoughts, which, therefore, can be said to employ indexical self-reference and be immune to error through misidentification relative to the concept "I".

Part II opens with an article by Benjamin Young (University of Nevada) that will try to demonstrate that the theoretical framework of Enactivism cannot account for olfactory perception. The main argument will show that the motoric component of olfaction – one of the central ideas of Enactivism – is not a necessary condition for perceiving smells, undermining the main intuition of the theory.

Next we will have an article by Bryan Frances (Lingnan University) discussing the difficulty of having a true ontology (physicalist or other) of the afterimages – the concept that refers to a visual experience that appears in one's vision after the exposure of the original image disappears. The author will then discuss four hypotheses: afterimages don't exist; they exist as an external physical thing; they exist as an internal physical thing; or they exist as a non-physical thing. It will be shown that there is a big difficulty to accommodate this phenomenon in a plausible ontology.

The following article by William Child (University of Oxford) will consider the causal theory of vision from philosophical and psychological perspectives. The first step is to consider the common idea that perceptual experience causally explained is based on a naïve, pre-theoretical thesis on vision rather than a scientific based explanation. The author's two main

goals are to consider the objection that the causal thesis cannot be part of the folk concept of vision and then, based on experimental work, discuss the causal theory of vision in the light of psychological work on causal understanding.

Lastly, Part II will end with an article by Keyvan Yahya (Chemnitz University of Technology) that will address how the influence of computational modelling of selective attention has been causing progress on the functional task of visual template identification.

Part III will start with an article by Steven S. Gouveia (University of Minho) and Diana Neiva (University of Porto) dealing with a new issue that may guide the future of Artificial Intelligence. The problem of the Mind-uploading has been debated in various disciplines and seems to raise old issues in the Philosophy of Mind: what is the nature of consciousness? Can Artificial Intelligence create artificial minds? It will also be discussing the various theories and their answers to the problem, proposing an alternative that seeks to break with a traditional conception of AI.

Next, Eray Özkural (Bilkent University) will discuss one of the most important issues of today's world: the ethics of (the future of) Artificial Intelligence. It will be shown that the idea of an existential risk to mankind is a scientifically implausibility, concluding with the suggestion that a beneficial AI agent with intelligence beyond human-level is possible and it will benefit the human society.

Finally, Matt Mahoney (Florida Institute of Technology) will present a relevant issue raised by Artificial Intelligence research: how much does it cost to have an automating human labor worldwide? The author will try to answer the question taking into account the detailed costs of hardware and software. Finally, some questions concerning the ethics of an expensive AI will also be raised.

Part IV follows, starting with an article by Alfredo Pereira Jr. (São Paulo State University) that will discuss several hypotheses on cortical integration and possible links between cortical processes and perceptual integration – the so called "binding problem". The author will propose an analogic model that combines subcortical control of cortical activity with mechanisms intrinsic to the cortical tissue.

Following, João de Fernandes Teixeira (Federal University of São Carlos) will examine the rise of the Cognitive Neuroscience and how this new science seeks to replace several academic disciplines (Psychology and Philosophy) for a brain-based approach that will solve all the problems. The influence of the Neurocartography (the new version of a brain-based phrenology) will be analyzed, raising several difficulties that this view can have.

To close Part IV, Judite Zamith-Cruz (University of Minho) and André Zamith-Cruz (University of Liverpool) will discuss the understanding of consciousness from five paradigm shifts and several theories and beliefs that are shaping our knowledge of the human mind.

Part V will open with an article by Klaus Gärtner (University of Lisbon) who will discuss views on the Transparency Thesis and its relationship with the privileged access to conscious experience. The main idea of the author is to show how the former influences the discussion of the latter and how they can be brought together in a compatible way.

The following article, by Aleksandar Risteski (University of Novi Sad), will examine the cartesian dualism as a consequence of the reductionism of its epistemology. The core of the argument is to show that Descartes' dualism is not a metaphysical consequence, but a gnoseological one. It will be shown how this position raises several ambiguous problems.

Jaime Milheiro will follow discussing several thematic influences on some of his work (Psychoanalysis, Psychiatry and Psychology). The main focus will be on rejecting the reductive methodology that the new sciences tend to apply to the issues of the mind and the brain, and how that can be a huge error to solve the mystery of the mind.

Next, José Antunes will examine the influences of Eastern Philosophy – its diversity and originality – on some of the central concepts of Philosophy of Mind.

At last, Manuel Curado will present a controversy that happened in the late nineteenth century about the role of consciousness. The idea, advocated by Dr. José de Lacerda, states that consciousness will disappear as evolution progresses – consciousness is considered an imperfection. It will be shown how this debate is important to the contemporary philosophy of mind.

PART I:

THE SELF IN CONTEMPORARY PHILOSOPHY OF MIND

CHAPTER ONE

'WHERE AND I, OR WHAT?': TWO WAYS OF BEING UNABLE TO GO WRONG WHEN ENCOUNTERING ONESELF (AND WHAT WE CAN LEARN FROM THEM)

SOFIA MIGUENS

I. Encountering oneself

In the Conclusion of the first Book of his *Treatise of Human Nature,* David Hume asks (rather dramatically) "Where Am I, or what?" – let us call this question 'Hume's question'. I want to begin with some examples of very different ways we risk going wrong when we try to answer Hume's question.

If we consider the literature of philosophy of language, and philosophy of self-knowledge, we come across the well-known Mach, Castañeda or Perry cases, where A has thoughts about B without realizing that B is A, and he is A (**the shabby pedagogue**[1], the editor of Mind[2], the shopper at the supermarket[3]). If we consider cognition, we come across experimental paradigms such as the Rubber Hand Illusion, where I am certain that this hand I see in front of me is *my hand* and that I feel it being touched, yet it

[1] In a famous footnote to his book *The Analysis of Sensations* (Mach 1914), p. 4. Quoted and commented by John Perry in "Identity, Personal Identity and the Self", p. 192.
[2] Castañeda, "On the Phenomeno-logic of the I", in Cassam ed. Self-Knowledge, p. 161.
[3] Perry, "The Problem of the Essential Indexical", in Cassam ed. Self-Knowledge, p. 167.

is an artificial hand, a Rubber Hand[4]; or pathologies, such as schizophrenia, where patients with verbal hallucinations may report thoughts suddenly occurring in them (e.g. 'Kill God!') of which they cannot possibly be the author. If we prefer thought-experiments, there is G. Evans' *Varieties of Reference* scenario of the tampered brain-limbs connections, where I think and feel that *my legs are crossed* and I believe I cannot be wrong about this but then I learn that the wiring was messed up with, and my brain is getting the stimulation from B's legs, so it seems perfectly plausible that I feel legs that are not mine as mine, as 'me'.[5]

These are, as I said, quite disparate phenomena of lack of authority in self-identification, yet they have motivated my leading question in this article: if all of this is possible, is there any way one is simply unable to go wrong when *encountering oneself*?

I will compare two answers to this question: D. Davidson's and S. Gallagher's. Neither Davidson nor Gallagher follows Hume in looking inward and searching for (an elusive) self, when looking for Myself, then stumbling on nothing but perceptions. They do have that in common; yet their accounts of 'the subjective' (I am borrowing the term from Thomas Nagel) are remarkably dissimilar.

Davidson proposes a language-based account of the subjective as *first-person authority*, whereas Shaun Gallagher proposes a phenomenology-inspired account of *immunity to error through misidentification (IEM) in proprioception*. As we will see, their respective focuses on language and perception lead them in quite different directions when pursuing a view of the subjective. This is why I think it might be fruitful to play them against each other. So how do we go about answering Hume's question and not looking inside oneself for a 'self'?

II. Davidson's way: Meaning What We Say and Knowing What We Mean

In his articles on the subjective, collected in the 2001 volume *Subjective, Intersubjective Objective*, Donald Davidson defends that subjectivity is (nothing but) first-person authority. Once we get rid of the (Cartesian) idea of subjectivity as a 'parade of objects before the mind',

[4] When their left arm is placed out of sight and the real hand (out of sight) and the rubber hand are simultaneously stroked, subjects experience the rubber hand as theirs.

[5] Evans (1982), *The Varieties of Reference,* Reprinted in Cassam, p. 198.

objects such that they 'must be what they seem and seem what they are', all we are left with is privacy and asymmetry.

Note that Davidson rejects the idea that there is no difference in type between self-knowledge and knowledge of other minds; unlike, say, G. Ryle, he accepts the doctrine of privileged access. But he does so in a very deflationary tone.

This is how he puts things: in order to know what other people think, I do need evidence, and I do take as evidence what they say and do. How can it be that in my own case I don't have to appeal to any evidence in order to know what I think? His answer is that there is an assumption, built into the very nature of interpretation, according to which a speaker usually knows what he means, whereas there is no such presupposition in the interpretation of others. First-person authority is thus a necessary feature of the interpretation of speech.

So Davidson's approach to the subjective is (1) directed at self-knowledge as knowledge of one's own thoughts, (2) posed in terms of language and interpretation. It is as such that 'the subjective' is dealt with as a part of Davidson's programme, which is an inquiry into the very possibility of thought and objective knowledge.

In the "Myth of the Subjective", Davidson acknowledges that this phenomenon, first-person authority, may give rise to the idea of epistemic priority of thought to world, and thus to scepticism[6]. But his account is deflationary precisely in that Davidsonian first-person authority is both more basic and less significant than Cartesian epistemic authority. It is simply a matter of our condition as linguistic creatures. And this is a condition in which I know what I mean when I say what I mean but that is the 'deepest' access I have to what I think (to this it should be added that there is no transparency of content of my thoughts to myself: Davidson is a content externalist).

Before we take this to be insufficient as a view of the subjective, we should keep in mind that it is put forward as part of an extremely ambitious philosophical programme. The account of the 'the subjective' is part of an investigation into the very possibility of objective knowledge and thought in minds such as our own, which, building on a Tarskian theory of truth, and 'using' it as a theory of interpretation for natural languages (the so-called 'radical interpretation'), ends up in a proposal to relate the objective, the intersubjective and the subjective (the tripod, Davidson calls it) in conceiving the nature of thought-world relations.

[6] Davidson, *The Myth of the Subjective*, pp. 43, 45, 47.

Ultimately, the core claim is that subjective-intersubjective-objective come together: only when the tripod is in place can there be, for Davidson, such a thing as e.g. a belief (mine, yours) about the objective world that can be true, as opposed to passing glimpses in a mental life, which have no claim to objectivity or truth[7].

But even allowing for this ambition, the fact is that Davidson's approach to the subjective takes place within an interpretation theory, and an interpretation theory as such simply assumes that there is something out there to be interpreted. Its touchstone is behavioural evidence; ultimately *stimulae*. In other words, in spite of his criticism of Quine, a view from the outside, a priority of a third-person perspective is, we may say, still at the core of Davidson's philosophy, and thus also in the heart of his view of the subjective. This is not a contingent detail. It is connected with another problem of Davidson's approach: the apparent absence of what we may call a view of perception (there are *stimulae*, there is the web of belief, and that is all).

Leaving perception out of the picture Davidson leaves our being acquainted with one's own bodily being out of the picture. So even if Davidson's view of the 'subjective' as first-person authority gives us important ideas – above all separating first-person authority and epistemic privilege, but also the importance of subjective-intersubjective-objective 'tripod' in accounting for human linguistic thought – his version of one's encountering oneself leaves us with something like an 'isolation in language and by language', a a-wordly subjectivity. And so it may seem that there is something missing.

III. Reintroducing perspective and perception: Shaun Gallagher on IEM

If we are looking for an alternative approaches to the phenomenon of immunity to error through misidentification (IEM) which focus on proprioception should be considered. My example here will be Shaun Gallagher.

In a 2012 article, "Immunity to Error through misidentification and the first-person", he defends IEM in proprioception against claims of people such as John Campbell, Elizabeth Pacherie and Marc Jeannerod. Campbell

[7] This means, of course, that, according to Davidson, the fact that there are minds in possession of the concepts of belief and truth is a condition for the existence of objective thought.

proposed that experiences such as hallucinations, thought insertion and delusions of control in which schizophrenic subjects report that their body is under the control of other people or things are counterexamples to IEM[8]. Pacherie and Jeannerod (in their 2004 *Mind and Language* article "Agency, simulation and self-identification") consider an even wider range of examples and conclude that such exceptions are sufficient for rejecting the principle:

> In a nutshell then, the bad news for philosophers is that self-identification is, after all, a problem. In the domain of action and intention at least, there is no such thing as immunity to error through misidentification, whether for the self as object (sense of ownership) or for the self as subject (sense of agency). The mechanisms involved in self and other attribution may be reasonably reliable but they are not infallible. (Pacherie et Jeannerod, 141*) [In other words, IEM obtains only contingently.]*

Gallagher does acknowledge that exceptions to IEM are abundant in both clinical and experimental situations: misidentifications of oneself as oneself range from somatoparaphrenia[9], to the Rubber Hand illusion, to virtual whole body displacement phenomena, to 'inserted thoughts' of schizophrenic patients, etc. Yet he believes that the principle (IEM) stills holds of proprioception; exceptions can be accounted for, as we will see, by isolating a conception of the subjective as irreducible 'perspective'.

But first let us have a brief look at the history of the discussion of IEM. When Sydney Shoemaker, following Wittgenstein, first spoke of IEM, what he had in mind was the use of the first-person pronoun 'I'[10] and the self-ascription of mental experience. In *The Varieties of Reference* chapter on Self-identification, Evans explored the fact that the phenomenon seems to extend from self-attribution of mental experience to proprioception. At that time (the 1980's) he was formulating his position against Thomas Nagel's view of the subjective according to which I cannot possibly think of myself as something in the world, 'the world as it is anyway' (in Williams' expression). Nagel thinks we *cannot make sense* of our own perspective, as subjects, as being part of the objective world, we cannot successfully locate consciousness in the objectively represented world.

---

[8] Cf. Campbell (1999), "Schizophrenia, the space of reasons and thinking as a motor process". In *The Monist*, 82 (4): pp. 609–625)

[9] Somatoparaphrenia patients deny the ownership of a limb connected to their body (even if looking at it in the mirror they, as it were, reclaim possession of it).

[10] Shoemaker, "Self-reference and self-awareness", in Cassam 1994.

From this he concludes for the existence of a 'gulf between the objective and the subjective' and posits what he terms an essentially perspectival subjective reality. Evans rejects Nagel's conclusion, which he thinks simply 'pressuposes Idealism'.

Pace Nagel, Evans believes that «Our thoughts about ourselves are in no way hospitable to Cartesianism. If there is to be a division between the mental and the physical, it is a division which is spanned by the Ideas we have of ourselves. Our customary use of 'I' simply spans the gap between the mental and the physical» (Evans, 1982. 'Ideas' is Evans' term for any Conception of myself). In other words, there is something wrong with Wittgenstein's initial distinction between subjective and objective uses of 'I' in the *Blue and Brown Books* (the distinction worked like this: «If I experience a toothache, it would be nonsensical to say 'Someone has a toothache, is it me'? On the other hand, for example, looking in the mirror and seeing a sunburned arm, I might say 'I have a sunburn'. But it is possible that I see someone else's arm in the mirror and mistake it for my own, and in that sense I seem to be misidentifying myself [while *referring to myself*] as 'object'».

Evans' scenario I mentioned at the beginning is formulated in the wake of criticism of Wittgenstein's distinction: 'My legs are crossed' – they are mine and I feel them – they are my legs and they are crossed - I cannot be wrong – or can I? What if wires are messed up with, and my brain gets the stimulation from A's legs? Can I feel legs that are not mine as mine, as 'me'?

Gallager's 2012 article is prompted by Evan's challenge. Each one of the psychiatry, neurology and cognitive science cases he considers (somatoparaphrenia, thought-insertion, Rubber Hand illusion, Nasa robots whose mechanic 'hands' its manipulators or controllers come to regard as their own, etc.) is a case of mistakenly identifying a body (or body parts, or thoughts, or actions) other than my own as being mine, or being me, as well as not identifying my body (or body parts, or thoughts, or actions) as being mine. All of them may be seen as exceptions to IEM. These are ways I can go wrong in taking myself to be the experiencer of my experiences, the thinker of my thoughts, the author of my actions, the owner of my body.

How can IEM possibly hold if there are all these exceptions? Gallagher's answer is that we need to bring apart [what he calls] the senses of self-agency and self-ownership which are, in normal cases, indistinguishable. We may in fact distinguish them in our *sense of mine-ness or ipseity*. An involuntary movement of my body makes the distinction clear: if I'm pushed from behind, there's sense of ownership of

my movement but not sense of agency – it's a movement of me, yet I am not 'authoring' it.

Sense of ownership is, in Gallagher's phenomenology-inspired terminology, the 'pre-reflective experience that I am the one undergoing the experience'. In contrast, sense of agency is the pre-reflective experience that I am the one causing or generating movement.

In his closer look at the sense of 'mine-ness' Gallagher is explicitly committed to a phenomenological conception of experience as pre-reflective self-acquaintance (Gallagher and Zahavi, 2010). Such conception is his step number one for defending IEM against claim such as by Jeannerod and Pacherie.

Step number two is proposing that IEM should be kept as independent as possible from particular modes of access to self which are, indeed, fallible and subject to manipulation, as experimental cases and pathologies show. So he claims that there is only one aspect of experience which remains self-specific and retains the characteristics of IEM – what he calls *first-person perspective.* He means *first-person perspective only*, and not sense of ownership, nor sense of agency. *What is first-person perspective then*? Gallagher's answer is that it is the non-relative bodily framework that acts as the origin point of the perspective and which «in action and perception is manifested in the integration of the non-relative bodily framework and the egocentric spatial frame of reference»[11].

This and only this survives manipulations of sense of ownership and sense of agency: even in cases such as the Rubber Hand illusion, somatoparaphrenia or thought-insertion, there is first-person perspective and that first-person perspective is still mine. I am the subject to whom I refer when I claim 'This arm (connected to my body) is not mine', or 'These thoughts are not mine'. Such embodied first-person perspective is according to Gallagher part of every action and perception as experiences. It is more basic than the linguistic phenomena of self-identification and self-reference at stake in the Mach-Castañeda-Perry cases, and is not contingent.

IV. What can we learn (for pursuing a view of the subjective)?

Something like Gallagher's phenomenology-inspired account of the subjective should be brought up against Davidson's view, where the wordly-situatedness of the subjective is simply lost.

[11] Gallagher, 2012, p. 261.

But is this 'irreducible perspective that is part of every perception and action' all that we want from a view of the subjective? Following Gallagher we leave behind Davidson's ambition: the framework of problems of thought and knowledge in which he thinks a view of the subjective belongs. Davidson's worries are, ultimately, epistemological and metaphysical worries; whereas Gallagher's are mostly psychological and cognitive. If we simply replace Davidson's proposal with a proposal such as Gallagher's we do not get done the same work done by a view of the subjective. So I want to suggest that although Gallagher does indeed point at something which is missing in Davidson's account of one's encountering oneself, not everything is wrong there.

Let us grant that first-person perspective is irreducible in a way which an interpretation theory such as Davidson's, with its stimulae-language dualism, and its a-wordliness, cannot account for – a phenomenological approach simply seems to fare better here. Still, spatial, self-locating, perspective is not all that what we are after when we pursue a view of the subjective. Why? Because it is not by itself sufficient for answering questions regarding truth, thought and knowledge in which a view of the subjective is involved. In order to pursue such questions we have to come to terms not only with first-person perspective but also with the fact that 1) *we are all first-persons* 2) *in the world*. This is what Davidson is after in his essays on the subjective, the intersubjective and the objective. His way of thinking about the tripod may not be the best but he is right to place a view of the subjective within a search for the 'shared standards of truth and objectivity that the very possibility of thought demands'.

That said, Gallagher's move, bringing in the spatial character of perspective and its irreducible nature, may prove very important in not going e.g. Nagel's way in conceiving the subjective: Nagel thinks that we *cannot make sense* of our own perspective, as subjects, as being part of the objective world. He takes a step from that to what Naomi Eilan calls the Metaphysical Elusiveness Claim, according to which 'I' stands for something external to the empirical world, a metaphysical subject[12]. When people such as Evans, Campbell or Eilan see the task of relating spatial thought and objectivity as important, they see it as a way of not taking Nagel's step. We can indeed *make sense* of our own perspective, as subjects, as being part of the objective world and spelling this out should

[12] Eilan (2013), 'Intelligible Realism About Consciousness: A Response to Nagel's Paradox'. In *Ratio*, 26 (3).

be an important component in a view of the subjective. [As for where to start] I end with a quote by someone who would agree with this:

> Any thinker who has an idea of an objective spatial world – an idea of a world of objects and phenomena which can be perceived but which not dependent on being perceived for their existence – must be able to think of his perception of the world as being simultaneously due to his position in the world and to the condition of the world at that position. The very idea of a perceivable, objective, spatial, world brings with it the idea of the subject being in the world (…) the idea that there is an objective world and the idea that the subject is somewhere cannot be separated and where he is is given as what he can perceive. (Gareth Evans, *The Varieties of Reference*, p. 200.)

CHAPTER TWO

EMPIRICAL PERSPECTIVES FROM THE SELF-MODEL THEORY OF SUBJECTIVITY: A BRIEF SUMMARY WITH EXAMPLES[1]

THOMAS METZINGER

(TRANS. LUÍS PINTO DE SÁ)

The goal of this chapter is to give a brief summary of the "self-model theory of subjectivity" (SMT) that is addressed to scientifically minded readers who are not themselves professional philosophers but who are nevertheless interested in philosophical theories of self-consciousness.[2] To

---

[1] This text is a greatly expanded, updated and revised version of an article first published in 2008 in *Progress in Brain Research*. I am grateful to Jennifer M. Windt for a variety of critical comments and suggestions for improvement. For their diligent help in the final correction, I am grateful to Hannes Boelsen, Regina Fabry and Lisa Quadt.

[2] For a popular scientific exposition of the fundamental concepts, with many examples, see Thomas Metzinger: Der Ego Tunnel. Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsethik. München 2014. For a comprehensive presentation of this theory in English, see Thomas Metzinger: *Being No One. The Self-Model Theory of Subjectivity*. Cambridge, Mass. 2003. The shortest, freely accessible summary of the theory can be found in *Scholarpedia 2* (2007), Art. 4174. A very accessible German overview of the theory was published in 2005: Thomas Metzinger: Die Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung in sechs Schritten. In: Christoph. S. Herrmann et al. (Eds.): *Bewusstsein. Philosophie, Neurowissenschaften, Ethik*. Stuttgart 2005, pp. 242–269. For a somewhat more substantial overview, covering the main conceptual tools with additional references to the literature but without any reference to empirical data, see Being No One – Eine sehr kurze deutsche

that effect, I will use a series of empirical examples from a number of different disciplines to illustrate some core ideas and to demonstrate the explanatory scope as well as the predictive power of SMT. The SMT is a philosophical and neuroscientific theory about what it means to be a self. It is also a theory about what it means to say that mental states are "subjective" and that a certain system has a "phenomenal first-person perspective." One of this theory's ontological claims is that the self is not a substance in the technical and philosophical sense of something that could "keep itself in existence", even if the body, the brain, or everything else in the physical universe disappeared. It is not an ontologically autonomous, self-subsistent entity, an individual or mysterious *Something* in the metaphysical sense. On that account, no such things as selves exist in the world: selves and subjects are not part of the irreducible, enduring constituents of reality.[3] What does exist is the *experience* of being a self,

---

Zusammenfassung. In: Thomas Metzinger: *Grundkurs Philosophie des Geistes. Bd. 1: Phänomenales Bewusstsein*. Paderborn 2006, pp. 424-475.

[3] For an overview of the various ways in which one can deny the existence of an ontologically autonomous "self" on philosophical-conceptual grounds, see Thomas Metzinger: The No-Self-alternative. In: Shaun Gallagher (Ed.): *The Oxford Handbook of the Self*. Oxford 2011, pp. 279-296. From an empirical perspective the following is clear: human beings are dynamic, socially-situated *systems.* Self-consciousness is a complex process which gradually produces certain skills that are conceptually best described as properties of a global system (rational thinking, selective attention, flexible and context-sensitive action control, linguistic self-reference, etc.). Many theoretical problems arise simply from the fact that these skills and global system properties are described incorrectly and thereby "reified". It is therefore perhaps important for non-philosophers to note that the enduring and widespread talk of "the I" or "my self" in folk psychology, media, but also in some academic contexts, constitutes a serious *logical* mistake. The personal pronouns of the first person singular – the linguistic expression "I" – always refer to the speaker who at that very moment employs it. Its logical function is not a generic concept or a reference to a concrete individual thing, but the self-localization of a speaker in a context of utterance. From a grammatical and semantical point of view, the "I" is also a singular term, which is tied to a specific context of utterance: that of the current speaker who employs a linguistic tool to point to themselves. In linguistic self-reference we nevertheless very often employ the indexical term "I" as if it were a name for an inner thing or a form of objectual reference, i.e. reference to an object (Maxwell R. Bennett, Peter M. S. Hacker: The philosophical foundations of neuroscience. Darmstadt 2010). But there is no special genus of things ("egos" or "selves") that one could carry in oneself, like a heart, or that one could possess, like a bicycle or a football. In addition, the ubiquitous talk of "our" or "my" self in everyday contexts is logically contradictory, since there will already have to be

as well as the diverse and ever changing contents of self-consciousness. This is what philosophers mean when they talk about the "*phenomenal* self": the way you *appear* to yourself, subjectively and at the level of conscious experience. The concept of the *phenomenal* self must therefore be sharply distinguished from the *substantial* self. The latter, as we have just seen, does not exist. In what follows, we shall always be referring to the phenomenal self.

Under SMT, this conscious experience of being a self is analyzed as the result of complex and dynamic self-organizing information-processing mechanisms and representational processes in the central nervous system. The phenomenal self is therefore not a substantial *thing*, but rather a discontinuous process. Of course, there are also higher-order, conceptually mediated forms of phenomenal self-consciousness that not only have neuronal, but also *social* correlates.[4] This theory, however, begins by focusing on the minimal representational and functional properties that a naturally evolved information-processing system – such as the *Homo sapiens* – has to have in order to later satisfy the constraints for realizing these higher order forms of self-consciousness. As most philosophers today would agree, the real problem lies in first understanding the simplest and most elementary form of our target phenomenon. This is the non-conceptual, pre-reflective and pre-linguistic layer in self-consciousness.

---

someone who "has" this self, i.e. a self beyond the self, which is related to this in a possession relation. The self cannot also be "in me", since then the very thing to which I would be identical would also be a proper part of mine.

[4] In this case, the SMT is often also a *person model*, and therefore the mental representation of an autonomous, rational subject. We experience then not merely as intelligent organisms, but, for example, as rational, ethical-integrity striving people. If we want to take such high-level human properties – rationality, morality or personality – really seriously, we need to investigate the gradual genesis of the very specific subpersonal functional profile which enables the self-organizing dynamics of social relations of recognition in the first place, through which these new properties come to be. For a more thorough discussion of the relationship between the conceptual and the non-conceptual content of self-consciousness, see Metzinger [2003]. Thomas Metzinger: Phänomenale Transparenz und kognitive Selbstbezugnahme. In: Ulrike Haas-Spohn [Ed.]: Intentionalität zwischen Subjektivität und Weltbezug. Paderborn 2003, pp. 411–459 is an earlier German version of this text. An hypothesis about the role of the unconscious self-model in the development of conceptually unmediated forms of social cognition can be found in Thomas Metzinger, Vittorio Gallese: The Emergence of a shared action ontology: Building blocks for a theory. In: Consciousness and Cognition 12 (2003), pp. 549-571.

Therefore, the first question we will have to answer is this: what are (relative to a class of systems, i.e. *Homo sapiens* or a particular kind of futuristic robot) the minimally sufficient conditions for the emergence of a conscious self? One could also subsequently ask what the *necessary* and sufficient conditions for all conceivable systems might be, but to answer this question is not the goal of the present text.

The self-model theory takes it that the properties in question are representational and functional brain properties. In other words, the psychological property that allows us to become a person in the first place is analyzed with the help of concepts from *sub-personal* levels of description. In philosophy of mind, this type of approach is sometimes called a "strategy of naturalization": a complex and opaque phenomenon – such as the emergence of phenomenal consciousness and a subjective, inward perspective – is conceptually analyzed in such a way as to make it empirically tractable. By reformulating classical problems from their own discipline, naturalist philosophers try to open them up for interdisciplinary investigations and scientific research programs, for instance in the cognitive and neurosciences. The American philosopher Josh Weisberg coined the expression "*method of interdisciplinary constraint satisfaction*" (MICS).[5] The method must simultaneously meet a variety of different levels of description, with both empirical and conceptual constraints, with an eye towards arriving at a comprehensive theory of self-consciousness. The hope is to arrive at a complex body of knowledge by a process of "triangulation", i.e. by making simultaneous use of various methods and sources of information in order to construct initially plausible and heuristically fruitful working concepts. These can then be refined and used to formulate testable hypotheses. It is a central task of the philosophy of cognitive science to develop adequate conceptual tools out of a metatheoretical perspective, tools that will enable the *integration* of the various levels of analysis and provide a formal framework which, ideally, can then merge different data sets and different theoretical approaches. SMT is an example of such an attempt.

A final introductory remark: the MICS, naturalism, and the search for a reductive account of the phenomenal self are not motivated by a scientistic ideology; instead, they are simply part of a rational research strategy. For

[5] Josh Weisberg: Consciousness Constrained – A Commentary on Being No One. In: PSYCHE. *An Interdisciplinary Journal of Research on Consciousness* 12 (2006).

instance, if it should turn out – as many people believe[6] – that there is something about human self-consciousness that lies *in principle* outside the reach of the natural sciences, then serious naturalistic philosophers would be satisfied with this finding as well. They would have achieved exactly what they set out to do in the first place: they would now have what philosophers like to call "epistemic progress." This type of progress could mean being able to describe, in a much more precise and fine-grained manner and with a historically unprecedented degree of conceptual clarity, *why exactly* is science unable to provide satisfying answers to certain questions, even in principle. Therefore, the most serious and respectable philosophical anti-naturalists will typically also be the ones who show the deepest interest in recent empirical findings. Naturalism and reductionism are not ideologies or potential new substitutes for religion – on the contrary, it is precisely the anti-naturalist and the anti-reductionist, who believe in the existence of an irreducible, essentially subjective element of the human mind, who will have the strongest ambition to make their philosophical case convincingly, in an empirically informed way, while precisely identifying the crucial points.

Step One: What Exactly Is the Problem?

What we erroneously call "the self" in folk-psychological contexts is the phenomenal self: that aspect of self-consciousness that is immediately given in subjective experience as the content of phenomenal experience. The phenomenal self may well be the most interesting form of phenomenal content. It endows our phenomenal space with two particularly fascinating *structural* features: centeredness and perspectivalness. As long as a phenomenal self exists, our consciousness is centered and bound to what philosophers call a "first-person perspective" (1PP). States inside this center of consciousness are experienced as *my own* states, because they are endowed with a sense of ownership that is prior to language or conceptual thought. In all of my conscious experiences and actions, I engage in constantly changing relations with the environment and with my own mental states. I experience myself as being *directed* – towards perceptual

[6] See for instance Thomas Nagel, *The View From Nowhere*, Oxford University Press, 1986, especially Chapter 4, which is also discussed in Thomas Metzinger: Perspektivische Fakten? Die Naturalisierung des „Blick von nirgendwo". In: Georg Meggle u.a. (Ed.): Analyomen 2. Perspektiven der Analytischen Philosophie. Berlin, New York 1997, pp. 103–110, and Metzinger (footnote 4).

objects, other human beings, or the contents of my own mental states and concepts. This process gives rise to a subjective inner perspective. The fact that I have such an inner perspective is, in turn, cognitively available to me.[7] In other words, what probably distinguishes human beings from most other animals is that we not only *have* a subjectively experienced inner perspective, but that we can also consciously conceptualize ourselves as *beings that have such an inner perspective*. We can attribute this property to ourselves conceptually and linguistically, for example, by applying the concept of a "subject" to ourselves.

The first problem, however, is that we are not exactly sure what we mean when we talk about these questions in this way. It is not just that we are not in a position to define with precision concepts like "I", "self", or "subject". The real problem is that these concepts often do not seem to refer to observable objects in the world. Therefore, the first thing we have to understand is how certain structural features of our inner experience determine the way we *use* these concepts. In order to analyze the logic of ascribing psychological properties to ourselves and to understand what these concepts actually refer to, we must first investigate the deep representational structure of conscious experience itself. Three higher order phenomenal properties are particularly interesting in this context:

- "*Mineness*": this is a higher order property of particular forms of phenomenal content. It is an immediately given, non-conceptual sense of ownership. Here are some examples of how we try to refer to this phenomenal property in folk-psychological discourse, using everyday language: "subjectively, *my* leg is always experienced as being a part of *me*"; "*my* thoughts and feelings are always experienced as part of *my own* consciousness"; "my volitional acts are always initiated *by myself*".

[7] For an introduction to the problem of cognitive self-reference as a potential difficulty for philosophical naturalism, see Lynne Rudder Baker: The first-person perspective: A test for naturalism. In: *American Philosophical Quarterly* 35 (1998). See also Metzinger (2003a) (Section 6.4.4) and especially Thomas Metzinger: Phenomenal transparency and cognitive self-reference. In: *Phenomenology and the Cognitive Sciences* 2 (2003), pp. 353–393. For an interesting and lucid criticism of my own account of the cognitive first-person perspective, see Lynne Rudder Baker: Naturalism and the first-person perspective. In: Georg Gasser (Ed.): *How successful is Naturalism?* Frankfurt am Main 2008 (Publications of the Austrian Ludwig Wittgenstein Society; 4), pp. 203–226.