PHILOSOPHY of Mind

Contemporary Perspectives

Edited by Manuel Curado and Steven S. Gouveia

Copyright 2017 Cambridge Scholars Publishing. All rights reserved. May not be reproduced in any form without permission from the publishers, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 2/14/2023 11:19 AM via
AN: 1307958 ; Manuel Curado, Editor, Steven S. Gouveia, Editor.; Philosophy of Mind: Contemporary Perspectives
Account: ns335141

Philosophy of Mind

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Philosophy of Mind:

Contemporary Perspectives

Edited by

Manuel Curado and Steven S. Gouveia

Cambridge
Scholars
Publishing

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

Copyright © 2017 by Manuel Curado, Steven S. Gouveia and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-5275-0000-4
ISBN (13): 978-1-5275-0000-6

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

TABLE OF CONTENTS

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

LIST OF ILLUSTRATIONS

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART I:

THE SELF IN CONTEMPORARY PHILOSOPHY OF MIND

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

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.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

- "*Selfhood*": this experientially untranscendable feeling of *being* a self is the essence, the phenomenal core property we are looking for. While “mineness” is concerned with part-whole relations, here we have a form of *global* identification, namely, of identification with the body or with the person as a whole. A few brief examples can again illustrate how we refer to this highly salient feature of our inner experience from the outside, using linguistic tools: "I am *someone*"; "I experience myself as *identical* across time"; "the contents of my self-consciousness form a coherent *whole*"; "without having the need to engage in any prior cognitive and reflexive operations I am *always* intimately familiar with the contents of my self-consciousness”.
- "*Perspectivalness*": in the context discussed here, perspectivalness is the dominant structural feature of phenomenal space as a whole: it is centered in an acting and experiencing subject, a self that engages in constantly changing relationships with itself and the world. Examples include: "my world has a fixed center, and *I* am this center"; "being conscious means having an *individual first-person perspective*"; "in experiencing persons and objects in the world as well as my own mental states, I am always bound to this inward perspective – I am its *origin*”.

The next step consists in a representational and functional analysis of these target properties. We must ask: what functional and representational properties does an information-processing system have to have in order to instantiate the *phenomenal* property in question? Which of these properties are sufficient, and are any of them strictly *necessary*? What *exactly* does it mean for such a system to experience the world as well as its own mental states from a first-person perspective? What we need is a consistent conceptual background that is sufficiently flexible to continually integrate new empirical findings and at the same time capable of taking the richness, the heterogeneity and the subtlety of phenomenal experience into account. I will now attempt to sketch the outline of such a conceptual framework in the remaining five steps.

Step Two: the Self-model

Step two consists in the introduction of a new theoretical entity: the phenomenal self-model (PSM). It is the most important part of the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

representational basis for instantiating the relevant phenomenal properties.[8] What is a mental "representation"? A representational state, for instance in the brain, is a state that has a certain *content*, because it is directed at something in the world. The brain-state is the physical carrier; the content is the meaning of this state. An inner representation is *about* something, it possesses semantic properties – having a correct representation implies *reference*. A representational state often functions as a placeholder for something external, the referent; it represents because it "stands" for something else. This "something" (what philosophers call an “intentional object”), however, can also be a past event, a potential future outcome, or even a mere possibility – in such cases, we speak of representations as *simulations*. They simulate *merely possible* states of affairs; they represent a possibility, not an actuality. SMT is predominantly a representational theory of consciousness, because it analyzes conscious states as representational states and conscious contents as representational contents.

One of our key questions was: which set of minimally sufficient *representational* properties does a system have to develop in order to possess the relevant target properties? This is our first, preliminary answer: the system needs a coherent self-representation, a consistent internal model of itself *as a whole*. In our case, the self-model is an episodically active representational entity whose content is determined by the system’s very own properties. Whenever such a self-representation is needed to regulate the system’s interactions with the environment, it is transiently activated – for instance in the morning, when we wake up. According to SMT, what happens when you wake up in the morning – when *you first come to yourself* – is that the biological organism, which you are, boots up its PSM: it activates the conscious self-model. This creates a whole bunch of new functional properties. First, the system can now for the first time focus its attention and some of its other cognitive abilities on itself *as a whole*. This also enables global forms of behavioral control because the awakened body can now for the first time causally control itself *as a whole*. Secondly, the system now *knows* that it has regained these functional properties, as the PSM makes this information available globally, and so also at the level of conscious experience, where they are presented to the PSM as *its own properties*.

[8] Robert Cummins: The Nature of Psychological Explanation. Cambridge 1983.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

In other words, what we need is a comprehensive theory of the self-model of the *Homo sapiens*.[9] Personally, I take it that this will be a predominately neurocomputational theory.[10] This means that there is not only a true representational and functional description of the human self-model, but also a true neurobiological description – for instance in terms of being a widely distributed, complex activation pattern in the brain.[11] The PSM is exactly that part of the *mental* self-model that is currently embedded in an integrated structure of the highest order, the global model of the world, and therefore available for steering introspective awareness.[12] An important aspect of this idea is that certain parts of the self-model can be both unconscious and functionally active at the same time. This consideration is of course of crucial relevance for the so-called "psychosomatic" medicine: "psychosomatic interactions" are in fact causal interactions between conscious and unconscious partitions of the self-model; a traumatizing experience could first be represented at the level of the PSM and then further processed in the unconscious self-functional model, where it then, for example, makes a direct causal contribution to an immunosuppression or the development of somatoform disorders. The encoding of traumatic information in the unconscious self-model would be a process by which this information would be connected to an existing cognitive structure (e.g. autobiographical or emotional self-representation)

[9] The methodological core of psychology – insofar as I may venture this type of metatheoretical observation from my standpoint as a philosophical observer – can now be analyzed in a fresh and fruitful way. Psychology is *self-model research*. It is the scientific discipline that focuses on the representational content, the functional profile and the neurobiological realization of the human self-model, including its evolutionary history and its necessary social correlates.

[10] See for instance Paul M. Churchland: A Neurocomputational Perspective. Cambridge 1989; Jakob Howhy: Mind in Prediction. Oxford 2013; Karl Friston: The free-energy principle: A unified brain theory? In: Nature Reviews Neuroscience 11 (2010), pp. 127–138; Jakub Limanowski, Felix Blankenburg: Minimal selfmodels and the free energy principle. In: Frontiers in Human Neuroscience 7 (2013), Art. 547; Anil K. Seth: Interoceptive inference, emotion, and the embodied self. In: Trends in Cognitive Sciences 17 (2013), pp. 565–573.

[11] A classical example is António R. Damásio: The Feeling of What Happens. Body and Emotion in the Making of Consciousness. New York e.o. 1999.

[12] Jack Yates: The content of awareness is a model of the world. In: Psychological Review 92 (1985), S. 249–284; Bernard Baars: A Cognitive Theory of Consciousness. Cambridge 1988; for a detailed analysis of the criteria for distinguishing different degrees of consciousness, see Metzinger, 2003a, Chapter 3.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and thus fed to a permanent storage. Because this process is a physical process in a complex system, it can also result in unexpected causal effects. The PSM is a coherent multimodal structure that probably depends on a partially innate, "hard-wired" model of the system's spatial properties (more about this in the second example[13]). This type of analysis treats the self-conscious human being as a special type of information-processing system: the subjectively experienced content of the phenomenal self is the representational content of a currently active, dynamic data structure in the system's central nervous system.

Aside from the representational level of description, one can also develop a *functional* analysis of the self-model. Whereas representational states are individuated by their content, a functional state is conceptually characterized by its *causal role*: the causal relationships it bears to input states, output states, and other internal states. An active self-model can therefore be seen as a sub-personal functional state: a set of causal relations of varying complexity that may or may not be realized at a given point in time. Since this functional state is realized by a concrete neurobiological state, it plays a certain causal role for the system. For instance, it can be an element in an account of information-processing. The perspective of classic cognitive science can help illustrate this point: the self-model is a *transient computational module* that is episodically activated by the system in order to control its interactions with the environment. In other words, what happens when you wake up in the morning, i.e., when the system that you are "comes to itself", is that this transient computational module is activated – the moment of "waking up" is exactly the moment in which this new instrument of intelligent information-processing emerges in your brain. It does so because you now

[13] See also the fifth section of Brian O'Shaughnessy: Proprioception and the body image. In: José Luis Bermúdez u.a. (Ed.): The Body and the Self. Cambridge 1995 and his use of the concept of a "*long-term body image*"; and Thomas Metzinger: Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewußtseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation. Paderborn 1993. Thomas Metzinger: Niemand sein. In: Sybille Krämer (Ed.): Bewußtsein. Philosophische Beiträge. Frankfurt am Main 1996. Thomas Metzinger: Ich-Störungen als pathologische Formen mentaler Selbstmodellierung. In: Georg Northoff (Ed.): Neuropsychiatrie und Neurophilosophie. Paderborn 1997. Antonio R. Damásio: Descartes' Error. New York 1994. Damásio 1999. For a good place to start delving into the empirical literature, see Manos Tsakiris: The sense of body ownership. In: Shaun Gallagher (Ed.): The Oxford Handbook of the Self. Oxford 2011.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

need a conscious self-model in order to achieve sensorimotor integration, generate complex, flexible and adaptive behavior, and attend to and control your body as a whole. The conscious self-model also has a metarepresentational layer because it provides the organism with an explicit representation of its own *abilities*.

The development of ever more efficient self-models as a new form of "virtual organ" – and this point should not be overlooked – is also a precondition for the emergence of complex societies. Plastic and ever more complex self-models not only allowed somatosensory, perceptual and cognitive functions to be continuously optimized, but also made the development of social cognition and cooperative behavior possible. The most prominent example, of course, is the human mirror neuron system, a part of our unconscious self-model that resonates with the self-models of other agents in the environment through a complex process of motor-emulation – of "*embodied simulation*," as Vittorio Gallese[14] aptly puts it – e.g., whenever we observe goal-directed behavior in our environment. Such mutually coupled self-models, in turn, are the fundamental representational resource for taking another person's perspective, for empathy and the sense of responsibility, but also for metacognitive achievements like the development of a *concept* of self and a *theory of mind*.[15] The obvious fact that the development of our self-model has a long biological, evolutionary, and (a somewhat shorter) social history can now be accounted for by introducing a *teleofunctionalist background assumption*, as it is often called in philosophy of mind.[16] The development

[14] Vittorio Gallese: Embodied simulation: From neurons to phenomenal experience. In: Phenomenology and the Cognitive Sciences 4 (2005), pp. 23–38.

[15] See, for instance, Doris Bischof-Köhler: Ichbewußtsein und Zeitvergegenwärtigung. Zur Phylogenese spezifisch menschlicher Erkenntnisformen. In: Annette Barkhaus et al (Ed.): Identität, Leiblichkeit, Normativität. Neue Horizonte anthropologischen Denkens. Frankfurt am Main 1996 and Doris Bischof-Köhler: Spiegelbild und Empathie. Bern 1993; on the possible neurobiological correlates of these basic social skills, which fit very well into the framework sketched above, see Vittorio Gallese, Alvin Goldman: Mirror neurons and the simulation theory of mind-reading. In: Trends in Cognitive Sciences 2 (1998), pp. 493–501 and Metzinger, Gallese (footnote 4); also Metzinger 2014, Chapter 6 (footnote 1).

[16] See, for instance, Ruth Garrett Millikan: Language, Thought, and other Biological Categories. Cambridge 1984; Ruth Garrett Millikan: White Queen Psychology and Other Essays for Alice. Cambridge 1993; Peter Bieri: Evolution, Erkenntnis und Kognition. In: Wilhelm Lütterfelds (Ed.): Transzendentale oder Evolutionäre Erkenntnistheorie? Darmstadt 1987; Fred Dretske: Explaining Behavior. Reasons in a World of Causes. Cambridge 1988; Daniel C. Dennett: The Intentional Stance.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and activation of this computational module plays a role *for* the system: the functional self-model possesses a true evolutionary description, i.e. it was a weapon that was invented and continuously optimized in the course of a "cognitive arms race".[17] The functional basis for instantiating the phenomenal first-person perspective can be seen as a specific cognitive achievement: the ability to use a centered representational space and thereby the ability to model oneself as an epistemic agent (see the last section of this article). In other words, phenomenal subjectivity (the development of a sub-symbolic, non-conceptual first-person perspective) is a property that is only instantiated when the respective system activates a coherent self-model and integrates it into its global world-model.

The existence of a stable self-model allows for the development of what philosophers call the "perspectivalness of consciousness": the existence of a single, coherent, and temporally stable reality-model that is representationally centered in a single, coherent, and temporally stable phenomenal subject, a model of the system *in the act of experiencing* or *in the act of knowing*.[18] This structural feature of the global representational space then leads to the episodic instantiation of a temporally extended, non-conceptual first-person perspective. If this global representational property is lost, this also changes the phenomenology and leads to the emergence of different neuropsychological deficits or altered states of consciousness. Some readers may have the impression that all of this is extremely abstract. A self-model, however, is not at all abstract – it is entirely concrete. A first, now classic, example will help demonstrate what – among many other things – I actually mean by the term "self-model".

In a series of fascinating experiments, in which he used mirrors to induce synesthesia and kinesthetic illusions in phantom limbs, Indian neuropsychologist Vilayanur Ramachandran demonstrated the PSM's plasticity. [19] Phantom limbs are subjectively experienced limbs that

---

Cambridge 1987; Fred Dretske: Die Naturalisierung des Geistes. Paderborn 1998; William G. Lycan: Consciousness and Experience. Cambridge 1996. The main texts can be found in a German translation (along with additional references also suitable for non-philosophers) in Vols. 2 and 3 by Thomas Metzinger [2006] (footnote 1), modules L - 15 and I- 9 to I- 11.

[17] Andy Clark: Microcognition. Philosophy, Cognitive Science, and Parallel Distributed Processing. Cambridge 1989, p. 61.

[18] This expression was first employed by António Damásio. See Damásio 1999, pp. 168ff.

[19] See Vilayanur S. Ramachandran, Diane Rogers-Ramachandran: Synaesthesia in phantom limbs induced with mirrors. In: Proceedings of the Royal Society B 263

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

typically appear after the accidental loss of an arm or a hand or after surgical amputation. In some cases, for instance following a non-traumatic amputation performed by a surgeon, patients have the subjective impression of being able to control and move their phantom limb at will. The neurofunctional correlate of this phenomenal configuration could consist in the fact that motor commands, which are generated in the motor cortex, continue to be monitored by parts of the parietal lobe and – since there is no contradictory feedback from the amputated limb – are subsequently integrated into the part of the self-model that serves as a *motor emulator*.[20] In other cases, the subjective experience of being able to move and control the phantom limb is lost. These alternative configurations may result from pre-amputational paralysis following peripheral nerve damage or from prolonged loss of proprioceptive and kinesthetic "feedback" that could confirm the occurrence of movement. On the phenomenological level of description, this may result in a paralyzed phantom limb.

Ramachandran and his colleagues built a "virtual reality box" by vertically inserting a mirror in a cardboard box from which the lid had been removed (Fig.1 illustrates the basic principle). The patient, who had been suffering from a paralyzed phantom limb for many years, was then told to insert both his real arm and his phantom arm into two holes that had been cut in the front side of the box. Next, the patient was asked to observe his healthy hand in the mirror. On the visual input level, this generated the illusion of seeing both hands, even though he was actually only seeing the reflection of his healthy hand in the mirror. So, what happened to the content of the PSM when the patient was asked to execute symmetrical hand movements on both sides? This is how Ramachandran describes the typical outcome of the experiment:

(1996), pp. 377-386; a popular account can be found in Vilayanur S. Ramachandran, Sandra Blakeslee: Phantoms in the Brain. New York 1998, pp. 46ff. The figure was published courtesy of Ramachandran.

[20] Related ideas are discussed by Rick Grush: The architecture of representation. In: Philosophical Psychology 10 (1997), pp. 5–25 and Rick Grush: Wahrnehmung, Vorstellung, und die sensomotorische Schleife. In: Heinz-Dieter Heckmann, Frank Esken (Ed.): Bewußtsein und Repräsentation. Paderborn 1998, p. 174; see also Ramachandran, Rogers-Ramachandran (footnote 19), p. 378.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Fig. 2-1. Mirror-induced synesthesia. Making part of a hallucinated self available for conscious action control by installing a virtual source of visual feedback. (picture courtesy of Vilayanur Ramachandran).

I asked Philip to place his right hand on the right side of the mirror in the box and imagine that his left hand (the phantom) was on the left side. “I want you to move your right and left arm simultaneously”, I instructed.

"Oh, I can’t do that”, said Philip. "I can move my right arm but my left arm is frozen. Every morning, when I get up, I try to move my phantom because it’s in this funny position and I feel that moving it might help relieve the pain. But”, he said looking down at his invisible arm, "I never have been able to generate a flicker of movement in it.”

"Okay, Philip, but try anyway.”

Philip rotated his body, shifting his shoulder, to "insert" his lifeless phantom into the box. Then he put his right hand on the other side of the mirror and attempted to make synchronous movements. As he gazed into the mirror, he gasped and then cried out, "Oh, my God! Oh, my God, doctor! This is unbelievable. It’s mind-boggling!". He was jumping up and down like a kid. "My left arm is plugged in again. It’s as if I’m in the past. All these memories from years ago are flooding back into my mind. I can move my arm again. I can feel my elbow moving, my wrist moving. It’s all moving again.”

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> After he calmed down a little I said, "Okay, Philip, now close your eyes".
>
> "Oh, my", he said, clearly disappointed. "It's frozen again. I feel my right hand moving, but there's no movement in the phantom."
>
> "Open your eyes."[21]

By now, it should be clear how these experimental findings illustrate the concept of a "self-model" I introduced above; what is moving in this experiment *is* the PSM. What made possible the sudden occurrence of kinesthetic movement sensations in the lost sub-region of the self-model was the installation of an additional source of feedback, of "virtual information". This immediately created a new functional property, let us call it "availability for selective motor control". By providing access to the visual mode of self-simulation, this made the corresponding information available to volition as well. Volitional control was now once again possible. This experiment also shows how phenomenal properties are determined by computational and representational properties. Bodily self-consciousness is directly related to brain processes.

The next example also concerns the phenomenology of phantom limbs. How "unreal" or "ghostly" are phantom limbs? Can we measure the "subjective experienced "realness" of the conscious self? A recent case study by Peter Brugger and his colleagues at the University of Zürich introduced a vividness rating on a 7-point scale that showed highly consistent judgments across sessions for their subject AZ, a 44-year-old university-educated woman born without forearms and legs. For as long as she can remember, she has experienced mental images of forearms (including fingers) and legs (with feet and first and fifth toes).

[21] See Ramachandran 1998, p. 47f. For the clinical and experimental details, see Ramachandran and Rogers-Ramachandran, 1996.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

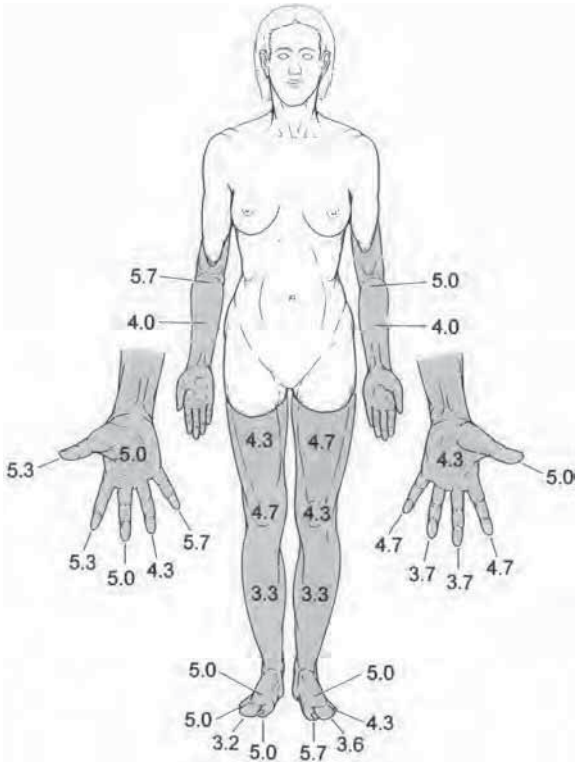

Fig. 2-2. *Evidence for an innate component of the PSM?* Phantom limbs (shaded areas) in a subject with limb amelia. The numbers are vividness ratings for the felt presence of different phantom body parts on a 7-point scale, from 0 (no awareness) to 6 (most vivid impression). (picture courtesy of Peter Brugger, Zürich).

The phantoms, as the figure above shows, were not as realistic as the content of her non-hallucinatory PSM. Functional magnetic resonance imaging of phantom hand movements showed no activation of the primary sensorimotor areas, but of the premotor and parietal cortex bilaterally. Transcranial magnetic stimulation (TMS) of the sensorimotor cortex consistently elicited phantom sensations in the contralateral fingers and hand. In addition, premotor and parietal stimulation evoked similar phantom sensations, albeit in the absence of motor-evoked potentials in the stump.

These data clearly demonstrate how body parts that were never physically developed can be phenomenally simulated in sensory and motor cortical areas. Are they components of an innate body model? Or could there be a more parsimonious explanatory hypothesis, which refers to the particular motivations of a severely disabled child and the existence of mirror neurons in brain regions such as BA 44? Could the completion of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the patient's body model perhaps be carried out through the visual observation during early childhood of other human beings moving around, through the arms and legs, so to speak, that via an unconscious form of motor perspective would have been "mirrored into" the patient's self-model? As I am a philosopher and not a neuropsychologist, I will refrain from further amateurish speculation at this point.

Recent results from research on pain experiences in phantom limbs point, however, to the potential existence of a genetically determined neuromatrix whose activation pattern may form the basis of these rigid parts of the self-model and of the almost invariant background of bodily self-experience (the so-called "phylomatrix of the body schema"[22]). Another interesting empirical result is that more than 20% of children born without an arm or a leg later develop the realistic conscious experience of having a phantom limb. In the context of phenomenal "realness" and in terms of the integration of the bodily self-model into the brain's conscious reality model as a whole, it may also be interesting to note that, in this case, "awareness of her phantom limbs is transiently disrupted only when some object or person invades her felt position or when she sees herself in a mirror".[23]

What do the phenomenologies of Ramachandran's and Brugger's subjects have in common? The transition from stump to phantom limb is seamless from the perspective of the quality of phenomenal "mineness";

[22] See Ronald Melzack: Phantom limbs, the self and the brain. The D.O. Hebb memorial lecture. In: Canadian Psychology 30 (1989), pp. 1–16 and Ronald Melzack: Evolution of the neuromatrix theory of pain. The Privithi Raj Lecture. Presented at the Third World Congress of World Institute of Pain, Barcelona 2004. Pain Pract 5 (2005), pp. 85–94; on the concept of a "neurosignature" in bodily self-consciousness, see Ronald Melzack: Phantom limbs. In: Scientific American 266 (1992), p. 93; an important study on phantom limbs following aplasia and early amputation is Ronald Melzack et al.: Phantom limbs in people with congenital limb deficiency or amputation in early childhood. In: Brain 120 (1997), pp. 1603–1620. See also Leonie Maria Hilti et al.: The desire for healthy limb amputation: Structural brain correlates and clinical features of xenomelia. In: Brain 136 (2013), pp. 318–329.

[23] See Peter Brugger et al.: Beyond re-membering: Phantom sensations of congenitally absent limbs. In: Proceedings of the National Academy of Science USA 97 (2000), pp. 6167–6172, here p. 6168. For further details concerning the phenomenological profile, see ibid.; for an interesting experimental follow-up study demonstrating the intactness of the phenomenal model of kinesthetic and postural limb properties, see Peter Brugger et al.: Hand movement observation in a person born without hands: Is body scheme innate? In: Journal of Neurology, Neurosurgery, and Psychiatry 70 (2001), p. 276.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

subjectively, they are both part of one and the same bodily self, because the quality of ownership is distributed evenly among them. There is no gap or sudden jump in the sense of ownership. The emergence of the bodily self-model is based on a sub-personal, automatic process of binding features together, of achieving coherence, which is subject to causal influence. But what exactly is it that is being experienced? What is the *content* of experience? In *De Anima*, Aristotle said that the soul is the form of the physical body, which perishes together with it at death (*On the Soul*, II: 412a, 412b–413a). According to Spinoza, the soul is the idea that the body develops of itself (*The Ethics*, II: 12 and 13). In more modern terms, we might say that an "idea" is simply a mental representation – more precisely a *self*-representation – and that the content of self-consciousness is the introspectively accessible part of this self-representation, namely the PSM postulated by the self-model theory. Gestalt properties – like body shape – are *global* properties of a perceptual object. Could the self-model then not be a neural mechanism to represent exactly such global properties, a new tool to acquire knowledge about the organism *as a whole*? In his dialogue *Meno*, Plato develops for the first time the philosophical thought that some of our ideas could be innate. And this is still an interesting question for today's neuroscience of self-consciousness: does the PSM possess an innate component? Is the conscious body image a kind of "fixed idea," anchored in an inborn and genetically predetermined nucleus?

As already mentioned, I do not wish to speculate at this point. I would rather draw attention to another matter, one which is often overlooked and which does not affect the phenomenology, but the *semantic properties* and the epistemic status of the body model in our brains. I endorse, at first without argument, the ontological background assumption that both a mind-independent external world and the physical body actually exist. The consciously experienced body model is then precisely the structure that has an intrinsic semantics from the outset: it is as if it were grounded from the beginning and functionally anchored in its referent. Its referent always exists, even when the physical self-model exists, and quite simply because the model is a physical *part* of the body itself – for example, a specific dynamic pattern of activation, a distributed sub-symbolic representation in the brain. Because at least the support of the *phenomenal* body model is always implemented internally, there is precisely here a guaranteed reference, and therefore an element of certainty.[24] In principle, of course, it

[24] All the cases in which this element of certainty is apparently lost are therefore theoretically interesting. That might just be the case if you experience yourself –

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

could turn out that all the contents of this model are misrepresentations, but the basic assumption of their existence – so that there is at least a physical basis for the experienced process – still stands. This relationship is important if one wants to understand the self-organization of meaningful states under naturalistic background assumptions about the evolutionary origin of intentional and semantic properties in biosystems: anyone who wants to provide a reductive explanation of such properties as mere functional properties must clarify how cognitive functions ultimately emerge from the interaction between perception and physical activity over a long history of dynamic environment interactions. Intentionality is then necessarily an embodied phenomenon: the semantic content of our world model unfolds gradually out of its biological anchoring. Motor primitives become semantic primitives. Through the targeted interaction with the environment the former gradually become, so to speak, "infected with meaning".[25] Let us now turn to example no. 3. It comes from a different scientific discipline altogether, namely from the fascinating new field of evolutionary robotics. It shows a number of further aspects that the conceptual framework of SMT, the self-model theory, predicts and seeks to explain. First, a self-model can be entirely unconscious; i.e., it can frequently be seen as the product of an automatic "*bottom-up*" process of *dynamical self-organization*; second, it is not a "thing" (or a model of a thing) at all, but is based on a continuous, ongoing modeling process;

---

in a dream, in asomatic OBEs – as a non-physical, purely mental self, Thomas Metzinger: Why are dreams interesting for philosophers? The Example of minimal phenomenal selfhood, plus an agenda for future research. In: Frontiers in Psychology 4 (2013), Art 746. It remains important here to distinguish precisely between the phenomenology of certainty and the epistemological aspect: there is a phenomenal core aspect of spatiotemporal self-localization, which may be the simplest form of self-consciousness, Olaf Blanke, Thomas Metzinger: Full-body illusions and minimal phenomenal selfhood. In: Trends in Cognitive Sciences 13 (2009), pp. 7-13; Jennifer M. Windt: Dreaming. Cambridge 2015. The dream example also shows that this aspect could be based on a misrepresentation that is not accessible to the subject as such (see Thomas Metzinger, Jennifer M. Windt: Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In: Thomas Grundmann u.a. (Ed.): Experimentelle Philosophie. Frankfurt am Main 2014; Metzinger 2013.

[25] This relationship makes it possible then to evolve a shared semantics for public representation systems in *groups* of biosystems through dynamically coupled self-models – a semantics for the gestures, vocalizations and linguistic symbols understandable by all group members. See Luc Steels, Manfred Hild: Language Grounding in Robots. Boston 2012.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

third, it can exhibit considerable *plasticity* (i.e., it can be modified through learning); and fourth, in its origins it is not based on language or conceptual thought, but very likely on an attempt to organize motor behavior. It is a computational tool to achieve global control and the structuring of the perceptual space. More precisely, a body-model has the function of integrating sensory impressions with motor output in a more intelligent and flexible manner. The unconscious precursor of the PSM clearly was a new form of intelligence and robustness.

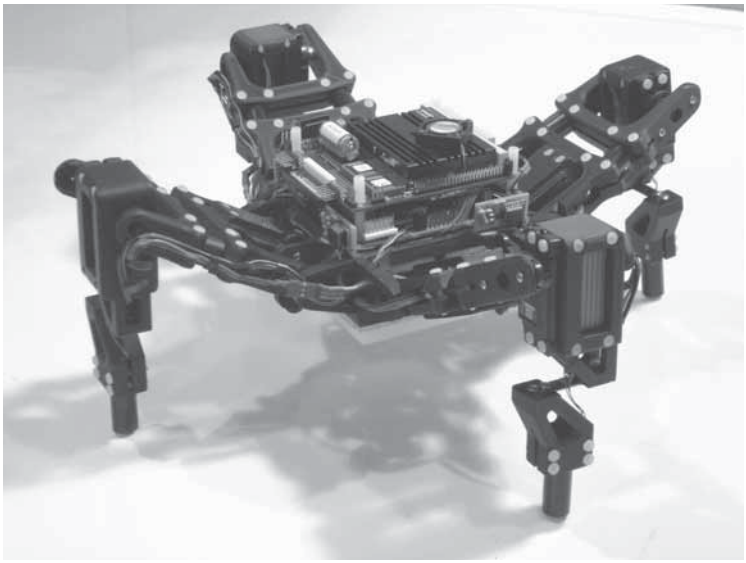

Fig. 2-3. Starfish, a four-legged physical robot, that walks by using an explicit internal self-model that it has autonomously developed and that it continuously optimizes. If it loses a limb, it can adapt its internal self-model. (photograph by Josh Bongard).

Bongard et al.[26] have created an artificial "starfish" that gradually develops an explicit internal self-model. This four-legged machine uses *actuation-sensation relationships*, i.e. relationships between self-generated body movements and feedback from these through actively altered sensory perception, to indirectly infer its own structure, and then uses this self-

[26] Josh Bongard et al.: Resilient machines through continuous self-modeling. In: Science 314 (2006), p. 1118.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

model to generate forward locomotion. When part of its leg is removed, it adapts its self-model and generates alternative gaits – it learns to limp.

In other words: unlike the phantom-limb patients presented in example no. 1 and no. 2 (and like most ordinary patients), the robot is able to *restructure* its body-representation following the loss of a limb. It can learn. This concept may not only help develop more robust machines and shed light on self-modeling in animals, but is also theoretically interesting, because it demonstrates for the first time that a physical system has the ability, as the authors put it, to "autonomously recover its own topology with little prior knowledge"[27] by constantly optimizing the parameters of its own resulting self-model. Starfish not only synthesizes an internal self-model, but also uses this self-model to generate intelligent behavior. The next figure gives an overview over this process **(Fig. 2-4).**

As can be seen, the robot initially performs an arbitrary motor action and records the resulting sensory data (one may also think, for example, of the random leg movement, the “motor babbling” of a human infant). The model synthesis component then synthesizes a set of 15 candidate self-models using stochastic optimization to explain the observed sensory-actuation relationship. The robot then synthesizes an exploratory motor action that causes maximum disagreement among the different predictions of these competing self-models. This action is carried out physically, and the 15 candidate self-models are subsequently improved using the new data. When the models converge, the most accurate model is used by the behavior synthesis component to create a desired behavior that can then be executed by the robot. If the robot detects unexpected sensor-motor patterns or an external signal resulting from unanticipated morphological change, it reinitiates the alternating cycle of modeling and exploratory actions to produce new models reflecting this change. The most accurate of these new models is then used to generate compensatory behavior and recover functionality.

[27] Ibid., p. 1120.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

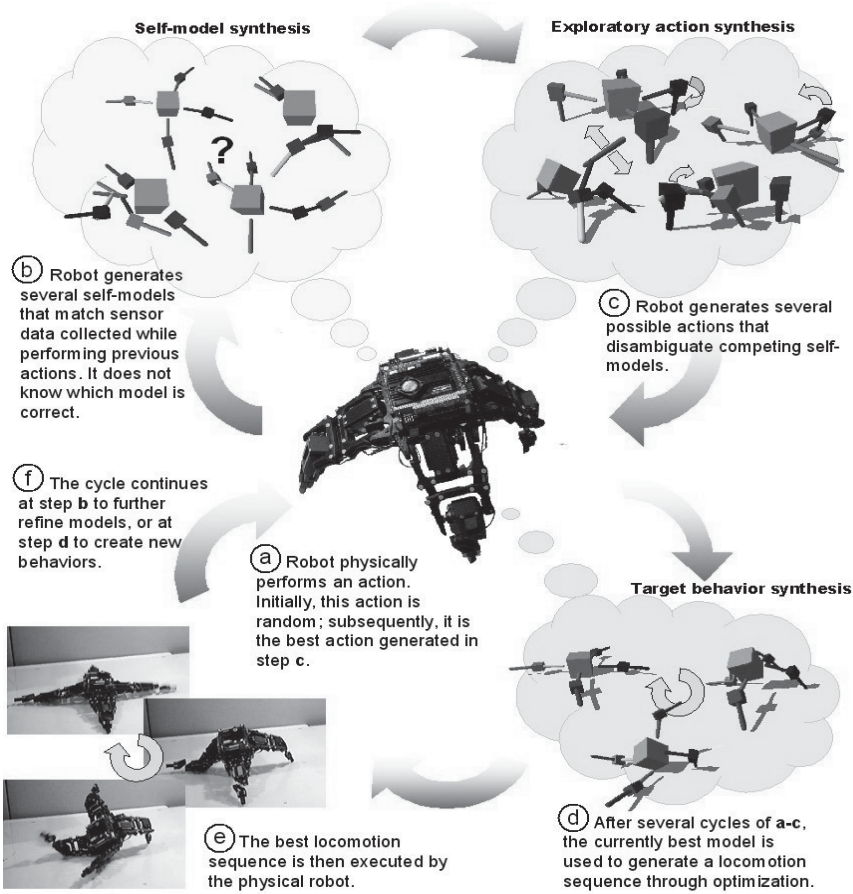

Fig. 2-4. (a and b) Self-model synthesis. The robot physically performs an action (a). Initially, this action is random; later, it is the best action found in (c). The robot then generates several self-models to match sensor data collected while performing previous actions (b). It does not know which model is correct. (c) Exploratory action synthesis. The robot generates several possible actions that disambiguate competing self-models. (d) Target behavior synthesis. After several cycles of (a)–(c), the best current model is used to generate locomotion sequences through optimization. The best locomotion sequence is executed by the physical device (e).

Technical details aside — what are the philosophical consequences of example no. 3? First, you do not have to be a living being in order to have a self-model. Non-biological SMT-systems are possible. Second, a self-

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

model can be entirely unconscious, i.e., it does not have to be a PSM, a *phenomenal* self-model. Consciousness is clearly a further step.[28] Third, a self-model supports planning and fast learning processes in a number of different ways. It clearly makes a system more intelligent and adaptable: its representational content is a *prediction* and is created in a systematic interplay between virtual and real behavior. Fourth, it is what I called above a virtual model or "virtual organ", and one of its major functions consists in appropriating a body by using a global morphological model to control it as a whole. Elsewhere, I have introduced the term "second-order embodiment" for this type of self-control (“third-order embodiment” refers then to the essential, more context-sensitive conscious process level).[29] If I may use a metaphor: one of the theoretical intuitions here is that a self-model allows a physical system to "enslave" its low-level dynamics with the help of a single, integrated and internal whole-system model, thereby controlling and functionally "owning" it, “appropriating”, so to speak, its own hardware on a causal level. We experience this appropriation subjectively as "mineness", and it is the decisive first step towards becoming an autonomous agent.

[28] See Thomas Metzinger (Ed.): Bewußtsein. Beiträge aus der Gegenwartsphilosophie. Paderborn 1995. For a first overview, see T.M.: The subjectivity of subjective experience: A representationalist analysis of the first-person perspective. In: T.M. (Ed.): Neural Correlates of Consciousness – Empirical and Conceptual Questions. Cambridge, MA: MIT Press 2000, pp. 285–306. See Metzinger, 2003a, Section 3.2 (footnote 1), for an additional list of ten constraints necessary for conscious experience.

[29] See Thomas Metzinger: Reply to Gallagher: Different conceptions of embodiment. In: PSYCHE. An Interdisciplinary Journal of Research on Consciousness 12 (2006); Thomas Metzinger: First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In: Lawrence Shapiro (Ed.): The Routledge Handbook of Embodied Cognition. London 2014. An important recent publication, which goes beyond the Starfish model and illustrates what I mean in philosophical and conceptual level by "second-order embodiment" is Malte Schilling, Holk Cruse: What’s next: Recruitment of a grounded predictive body model for planning a robot’s actions. Frontiers in Psychology 3 (2012), Art. 383.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Step Three: a Representationalist Analysis of the Three Target Properties

Here, the basic idea is that self-consciousness is, first of all, an *integrative process*: by becoming embedded in the currently active self-model, representational states acquire the higher order property of phenomenal mineness. If this integrative process is disturbed, this results in various neuropsychological syndromes or altered states of consciousness.[30] For example, one can analyze somatoparaphrenia[31], xenomelia[32] or certain positive symptoms of schizophrenia, such as thought insertion, as functional configurations in which the system can no longer integrate the existing representations of body parts or of their own cognitive processes into the PSM. Let us take a look at some examples of what happens when phenomenal mineness, the subjective sense of ownership, is selectively lost.

- Florid schizophrenia: "Consciously experienced thoughts are no longer *my* thoughts."
- Somatoparaphrenia, xenomelia (*body identity integrity disorder*, BIID)[33], unilateral hemi-neglect: "My leg is not *my* leg."
- Depersonalization[34] and delusions of control: "My body as a whole strikes me as foreign and unreal; I am a robot, I am turning into a puppet, and volitional acts are no longer *my* volitional acts."[35]

---

[30] For case studies, see Chapter 7 in Metzinger (2003a).

[31] Giuseppe Vallar, Roberta Ronchi: Somatoparaphrenia: A body delusion. A review of the neuropsychological literature. In: Experimental Brain Research 192 (2009), pp. 533–551.

[32] Melita J. Giummarra u.a.: Body Integrity Identity Disorder: Deranged Body Processing, Right Fronto-Parietal Dysfunction, and Phenomenological Experience of Body Incongruity. In: Neuropsychological Review 21 (2011), pp. 320–333, Hilti 2013.

[33] Giummarra et al.. 2011, Hilti 2013.

[34] Matthias Michal, Manfred E. Beutel: Weiterbildung CME: Depersonalisation/ Derealisation – Krankheitsbild, Diagnostik und Therapie. In: Zeitschrift für Psychosomatische Medizin und Psychotherapie 55 (2009), pp. 113–140.

[35] In this case, what philosopher and psychiatrist Karl Jaspers called *Vollzugsbewusstsein*, or "executive consciousness," is selectively lost; see Karl Jaspers: Allgemeine Psychopathologie. Berlin, Heidelberg [1946] 1973, p. 102. See also Daphne Simeon, Jeffrey Abugel: Feeling Unreal: Depersonalization Disorder and the Loss of the Self. Oxford, New York 2006, Matthew Ratcliffe:

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

- Manic disorders: “I am the whole world; all events in the world are controlled by *my own* volitional acts.”

Subjectively experienced "mineness" is a property of discrete forms of phenomenal content, such as the mental representation of a leg, a thought, or a volitional act. This property, the sense of ownership, is not necessarily connected to these mental representations, i.e., it is not an intrinsic, but a *relational* property. That a thought or a body part is consciously experienced as your own is not an essential, strictly necessary property of the conscious experience of this thought or body part. It could have been otherwise. In other phenomenological contexts, mineness disappears. Its distribution over the different elements of a conscious world-model can vary. If the system is no longer able to integrate certain discrete representational contents into its self-model, it is lost. If this analysis is correct, it should be possible, at least in principle, to operationalize this property by searching for an empirically testable metrics for the coherence of the self-model in the respective areas of interest. One could also empirically investigate how and in which brain areas a certain type of representational content is integrated into the self-model. Local body illusions like the rubber hand illusion [36], some dysfunctions in the voluntary movement system [37], and the phenomenon of hallucinated agency[38] appear as misrepresentations since they are already in the brain and their active representational content is embedded into the self-model, through which they are thereby automatically provided with the phenomenal property of "mineness": whatever is functionally embedded by the brain in the currently active PSM is inescapably experienced by the person concerned as *her own* state. Here is a concrete example of what I mean by "mineness”.

In the *rubber-hand illusion* (RHI), the sensation of being stroked with a probe is integrated with the corresponding visual perception in such a way that the brain transiently matches a proprioceptive map (of the

---

Feelings of Being. Phenomenology, Psychiatry and the Sense of Reality. Oxford, New York 2008 (International Perspectives in Philosophy and Psychiatry).

[36] Matthew Botvinick, Jonathan Cohen: Rubber hands “feel” touch that eyes see. In: Nature 391 (1998), p. 756.

[37] See Thomas Metzinger: Conscious volition and mental representation: Towards a more fine-grained analysis. In: N. Sebanz, W. Prinz (Ed.): Disorders of Volition. Cambridge, MA 2006, pp. 19–48.

[38] See Daniel M. Wegner, Thalia Wheatley: Apparent mental causation: Sources of the experience of will. In: American Psychologist 54 (1999), pp. 480–492.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

subject's own-body perception) with a visual map (of what the subject is currently seeing). At the same time, the feeling of "ownership" or phenomenal "mineness" is transferred to the rubber hand. The subject experiences the rubber hand as *her own* hand and feels the strokes *in* this hand. When asked to point to her concealed left hand, her arm movement will automatically swerve in the direction of the rubber hand.[39]

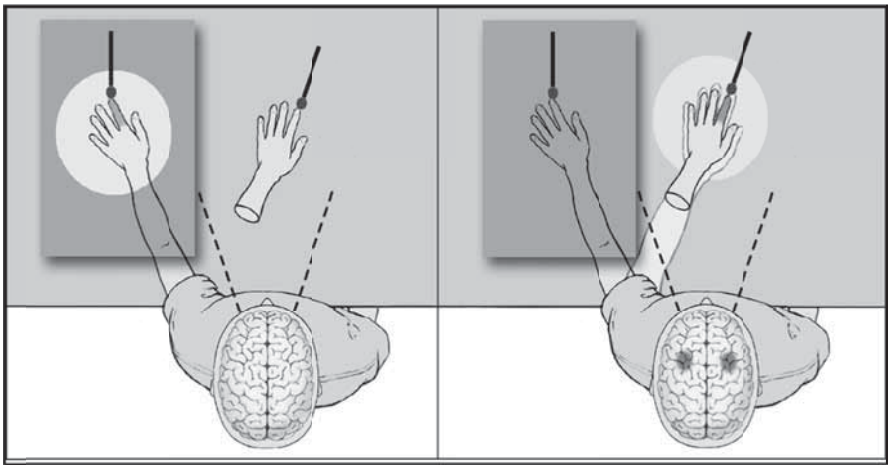

Fig. 2-5. *The rubber-hand illusion*. A healthy subject experiences an artificial limb as part of her own body. The subject observes a facsimile of a human hand while one of her own hands is concealed (grey square). Both the artificial rubber hand and the invisible hand are then stroked repeatedly and synchronously with a probe. The bright and dark areas indicate the respective tactile and visual receptive fields for neurons in the premotor cortex. The illustration on the right shows the subject's illusion as the felt strokes are brought into alignment with the seen strokes of the probe (the darker areas are those of heightened activity in the brain; the phenomenally experienced, illusory position of the arm is indicated by the bright outline). The respective activation of neurons in the premotor cortex is demonstrated by experimental data. (figure by Litwak Illustrations Studio, 2004).

If one of rubber hand's fingers is "hurt" by being bent backwards into a physiologically impossible position, the subject will also experience her real phenomenal finger as being bent much farther backwards than it is in reality. At the same time, this will also result in a clearly measurable skin

[39] Botvinick and Cohen, 1998, p. 756.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

conductance response. While only 2 out of 120 subjects reported an actual pain sensation, many subjects drew back their real hands, opened their eyes up widely in surprise, or laughed nervously.[40] Subjects also showed a noticeable reaction when the rubber hand was hit with a hammer. Again, it becomes clear how the phenomenal target property is directly determined by representational and functional brain processes. What we experience as part of our self depends on the respective context and on which information our brain integrates into our currently active self-model.[41] The intriguing question, of course, is this: could *whole-body* illusions exist as well? The answer is yes, and we will soon return to this point in example no. 5.

But first let us take a look at the second target property, at consciously experienced selfhood. Intuitively one might call this feeling "global mineness": the subjective feeling of possessing the body *as a whole*, and of phenomenally *identifying* oneself with it. This description, however, would lead us to conceptual problems because it introduces an invisible self "behind" the body, which owns the body. Better to say, then, that the body presents itself with two high level realities: that of its own existence as a totality and the ability to control such a totality causally. It comes down to the representation of existence and autonomy. Methodologically, it is important to first isolate the *simplest* form of the target.[42] Phenomenal selfhood corresponds to the existence of a single, coherent and temporally stable self-model that constitutes the center of the representational state as whole. If this representational module is damaged or disintegrates, or if multiple structures of this type alternate or are simultaneously activated by the system, this will again result in various neuropsychological disturbances or altered states of consciousness:

- Anosognosia and anosodiaphoria: loss of higher order insight into existing deficits, e.g., in cortically blind patients who deny that they are blind (*Anton's Syndrome*).

[40] K. Carrie Armel, Vilayanur S. Ramachandran: Projecting sensations to external objects: Evidence from skin conductance response. In: Proceedings of the Royal Society B 270 (2003), pp. 1499–1506, here p. 1503.
[41] See especially Botvinick and Cohen, 1998; and the neuroimaging study by Matthew Botvinick: Probing the neural basis of body ownership. In: Science 305 (2004), pp. 782–783; H. Henrik Ehrsson, Charles Spence, Richard E. Passingham: That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. In: Science 305 (2004), pp. 875–877.
[42] Blanke, Metzinger (footnote 24).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

- *Dissociative Identity Disorder* (DID): the system uses different and alternating self-models as a means of coping with extremely traumatic and socially inconsistent situations.[43]
- "Ich-Störungen" or identity disorders: a large class of psychiatric disturbances connected to altered forms of experiencing one's own *identity*. Schizophrenia is a classical example, as are Cotard syndrome, reduplicative paramnesia, or delusional misidentification.[44]

The existence of a stable self-model also almost always gives rise to the "perspectivalness of consciousness" in terms of transient subject–object relationships.[45] This structural feature of the global representational space leads to the episodic instantiation of a temporally extended and non-conceptual first-person perspective. It, too, can be lost.

- Complete depersonalization: loss of the phenomenal first-person perspective, accompanied by dysphoric states and functional deficits ("dreadful ego-dissolution").[46]
- Mystical experiences: selfless and non-centered global states, which are experienced, viz. described, as non-pathological and

[43] For the current diagnostic criteria for DID, see DSM-IV: 300.14.

[44] For a discussion on why identity disorders are interesting from a philosophical perspective, see Thomas Metzinger: Why are identity disorders interesting for philosophers? In: Thomas Schramme, Johannes Thome (Ed.): Philosophy and Psychiatry. Berlin 2004.

[45] See step 6 below; see also Nagel, 1986; Metzinger, 1993, 1995a, 2005a. As for examples of conscious reality models that are egocentric but aperspectival (like maybe akinetic mutism), see in particular Thomas Metzinger: Conscious volition and mental representation: Towards a more fine-grained analysis. In: Natalie Sebanz, Wolfgang Prinz (Hg.): Disorders of Volition. Cambridge, MA 2006.

[46] See Adolf Dittrich: Ätiologie-unabhängige Strukturen veränderter Wachbewußtseinszustände. Stuttgart 1985, Adolf Dittrich: Ätiologie-unabhängige Strukturen veränderter Wachbewußtseinszustände. Ergebnisse empirischer Untersuchungen über Halluzinogene I. und II. Ordnung, sensorische Deprivation, hypnagoge Zustände, hypnotische Verfahren sowie Reizüberflutung. Berlin 1996; Adolf Dittrich, Daniel Lamparter, Maja Maurer: 5D-ABZ. Fragebogen zur Erfassung Außergewöhnlicher Bewusstseinszustände. Zürich 2006; Erich Studerus, Alex Gamma, Franz X. Vollenweider: Psychometric evaluation of the altered states of consciousness rating scale (OAV). In: PLoS ONE 5 (2010), e12412.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

unthreatening ("oceanic boundary loss," "*The Great View from Nowhere*").[47]

In order to do justice to the richness and the diversity of different forms of human experience, one has to acknowledge the existence of certain non-perspectival and selfless forms of conscious experience. Phenomenologically, *non-subjective* consciousness – phenomenal experience that is not tied to a self or an individual first-person perspective – is not only a possibility, but a reality, even if we may find this idea inconceivable.[48] Particularly philosophically interesting in that regard are all those classes of states in which conscious people spontaneously use the first-person pronoun "I", as in, for example, the Cotard delusion, or in the case of prolonged spiritual experiences. The self-model theory provides the conceptual means to account for these special cases.[49]

Example no. 5 will demonstrate this principle in another domain. If we have the necessary conceptual tools, we can not only take the subtleties and the variability of human experience seriously but we can also develop new interdisciplinary research programs that penetrate into "taboo zones" and shed light on phenomena that in the past were the targets only of esoteric folklore and metaphysical ideologies. Could there be an integrated kind of bodily self-consciousness, be it of a mobile body fully available for volitional control or of a paralyzed body that is entirely a phenomenal confabulation – in short, a *hallucinated* and a *bodily* self at the same time? Is it conceivable that something like a full-body analog of the rubber-hand-illusion or a "globalized phantom-limb experience" – the experience of a *phantom body* – could emerge in a human subject? The answer is yes. There is a well-known class of phenomenal states in which the experiencing person undergoes the untranscendable and highly realistic conscious experience of leaving her physical body, usually in the form of an etheric double, and moving around outside of it. In other words, there is

[47] See the aforementioned articles.

[48] All those states of consciousness in which there is no phenomenal self are seen by human beings as "unimaginable" or "counter-intuitive", precisely because these cannot actively be simulated (since the simulation would necessarily produce the phenomenology of internal action, i.e. "cognitive agency") and because such states, once given, can be integrated into the autobiographical self-model only with great difficulty (because they are not a part of their *own* phenomenal biography). See Metzinger (footnote 3).

[49] For additional neurophenomenological case studies, see Metzinger, 2003a, Chapters 4 and 7.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

a class (or at least a strong cluster) of intimately related phenomenal models of reality that are classically characterized and defined by *a visual representation* of one's own body from a perceptually impossible, externalized third-person perspective (e.g., seeing oneself from above, lying on the bed or on the road) plus a *second representation* of one's own body, typically (but not in all cases) freely hovering or floating in space. This second body-model is the locus of the phenomenal self. It not only forms the "true" focus of one's phenomenal experience, but also functions as an integrated representation of all kinesthetic qualia and all non-visual forms of proprioception. This class of phenomenal states is called the "out-of-body experience" (OBE).[50] Elsewhere[51], I have argued that our traditional, folk-phenomenological concept of a "soul" may have its origins in accurate and sincere first-person reports about the experiential content of this specific class of neurophenomenological states.

OBEs frequently occur spontaneously while falling asleep, but also following severe accidents or during surgical operations. At present, it is not clear whether the concept of OBE possesses a clearly delineated set of necessary and sufficient conditions. Instead, it may turn out to be a cluster concept constituted by a whole range of diverging (and possibly overlapping) subsets of phenomenological constraints, each forming a set of sufficient, but not necessary, conditions. On the other hand, the OBE clearly is something like a phenomenological *prototype*. There is a common core to the phenomenon, as can be seen from the simple fact that many readers will have already heard about this type of experience in one way or another.

One can offer a representationalist analysis of OBEs by describing them as a class of deviant self-modeling processes. Phenomenological reports of "soul travel" would then be, for example, reports on the representational content of the PSM during such a deviant state of consciousness. A prototypical feature of this class of deviant PSM seems to be the coexistence of (a) a more or less veridical representation of the bodily self as seen from an external visual perspective, which does *not*, however, function as the center of the global model of reality, and (b) a second self-model, which largely integrates proprioceptive perceptions in

[50] A brief summary of scientific studies can be found in Metzinger [2014] (footnote 2), p. 135ff .

[51] Thomas Metzinger: Out-of-body experiences as the origin of the concept of a „soul". In: Mind and Matter 3 (2005), pp. 57–84. For further references see also Bigna Lenggenhager, Tej Tadi, Thomas Metzinger, Olaf Blanke: Video Ergo Sum: Manipulating bodily self-consciousness. In: Science 317 (2007), pp. 1096–1099.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

subjective experience – although, interestingly, weight, temperature and pain sensations are only integrated to a lesser degree – and which possesses special properties of shape and form that may or may not be veridical. Although only one of these models of the system presents itself as the subject of experience, i.e. presenting itself *as experienced*, both models are located within the same spatial frame of reference (that is why they are *out of-body*-experiences). This frame of reference is an egocentric frame of reference.

This is the most commonly reported state of consciousness in which two system models are active at the same time. Of course, only one of them will be the "place of identity", the place also where the phenomenology is to be found, what in philosophy would be called the “acting subject”. The other self-model – that of the physical body, lying on the bed or on the operating table as below – is, strictly speaking, not a self-model, because it does not function as source of the 1PP.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

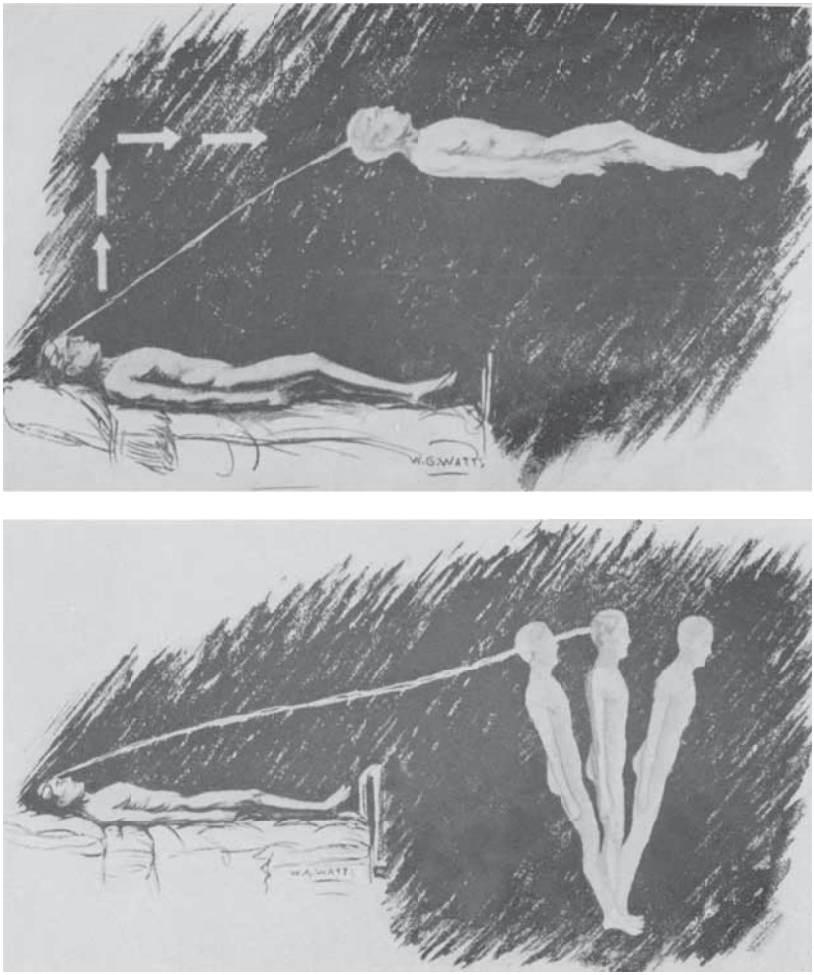

Fig. 2-6a. Kinematics of the PSM during an OBE-onset: the classical Muldoon-scheme. From Muldoon S. and Carrington, H. (1929). The Projection of the Astral Body. Rider & Co., London.

This second self-model is not the model of a *subject*. It is not, for instance, the place in the room from which you direct your attention at will onto the world. On the other hand, it is still your own body that you see down there from the outside. You recognize it visually as your own, but now it is no longer the body as a *subject*, as a place of knowledge, of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

action and conscious experience, the part of reality with which you *identify*. All of this makes the ego stand out. Such phenomenal configurations are therefore instructive, because they allow us to distinguish between different functional layers in the conscious self of human beings. Let us now look at two classical phenomenological descriptions of OBEs as spontaneously occurring in an ordinary non-pathological context.

> "I awoke at night – it must have been at about 3 a.m. – and realized that I was completely unable to move. I was absolutely certain I was not dreaming, as I was enjoying full consciousness. Filled with fear about my current condition, I had only one goal, namely to be able to move my body again. I concentrated all my will-power and tried to roll over to one side: something rolled, but not my body – something that was me, my whole consciousness including all of its sensations. I rolled onto the floor beside the bed. While this happened, I did not feel bodiless, but as if my body consisted of a substance in between the gaseous and the liquid state. To the present day, I have never forgotten the combination of amazement and great surprise that gripped me when I felt myself falling onto the floor, but without the expected thud. Had the movement actually unfolded in my normal physical body, my head would have had to collide with the edge of my bedside table. Lying on the floor, I was overcome by terrible fear and panic. I knew that I possessed a body, and I only had one great desire – to be able to control it again. With a sudden jolt, I regained control, without knowing how I managed to get back into it."

The prevalence of OBEs ranges from 8% in the general population to 25% in students, with extremely high incidences in certain subpopulations like, to name just one example, 42% in schizophrenics.[52] However, it would be false to assume that OBEs typically occur in people suffering from severe psychiatric disorders or neurological deficits. Quite the contrary, most OBE-reports come from ordinary people in everyday life situations. Let us stay therefore with non-pathological situations and look

[52] See Susan J. Blackmore: Spontaneous and deliberate OBEs: a questionnaire survey. In: Journal of the Society for Psychical Research 53 (1986), pp. 218–224; for an overview and further references see Carlos S. Alvarado: Research on spontaneous out-of-body experiences: A review of modern developments, 1960–1984. In: B. Shapin, L. Coly (Ed.): Current Trends in PSI Research. New York, S. 140–167; C.S.A.: Out-of-Body Experiences. In: Etzel Cardena, Steven J. Lynn, Stanley Krippner (Ed.): Varieties of Anomalous Experience. Examining the Scientific Evidence. Washington, D.C. 2000, pp. 138–218, here p. 18 und Metzinger [2014] (footnote 3), p. 124ff.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

at another paradigmatic example, again reported by Swiss biochemist Ernst Waelti:

> "I went to bed in a dazed state at 11 p.m. and tried to go to sleep. I was restless and turned over frequently, causing my wife to grumble briefly. Now, I forced myself to lie in bed motionless. For a while I dozed before feeling the need to pull up my hands, which were lying on the blanket, in order to bring them into a more comfortable position. At the same instant, I realized that I was absolutely unable to move and that my body was lying there in some kind of paralysis. Nevertheless, I was able to pull my hands out of my physical hands, as if the latter were just a stiff pair of gloves. The process of detachment started at the fingertips, in a way that could be clearly felt, almost with a perceptible sound, a kind of crackling. It was exactly the movement that I had actually intended to carry out with my physical hands. With this movement, I detached from my body and floated out of it head first. I moved into an upright position, as if I was almost weightless. Nevertheless, I had a body consisting of real limbs. You have certainly seen how elegantly a jellyfish moves through water. I could now move around with the same ease. I lay down horizontally in the air and floated across the bed, like a swimmer, who has pushed himself from the edge of a swimming pool. A delightful feeling of liberation arose within me. But soon, I was seized by the ancient fear common to all living creatures, the fear of losing my physical body. It sufficed to drive me back into my body." [53]

Sleep paralysis, of the kind occurring in the two examples cited above, is not a necessary precondition for OBEs. They frequently occur during extreme sports, in high-altitude climbers or marathon runners, for instance.

> A Scottish woman wrote that, when she was 32 years old, she had an OBE while training for a marathon. «After running approximately 12–13 miles I started to feel as if I wasn't looking through my eyes but from somewhere else. I felt as if something was leaving my body, and although I was still running along looking at the scenery, I was looking at myself running as well. My 'soul' or whatever, was floating somewhere above my body high enough up to see the tops of the trees and the small hills».[54]

[53] Ernst Waelti: *Der dritte Kreis des Wissens*. Interlaken 1983, p. 25; English translation by TM.
[54] See Alvarado, 2000, p. 184. English translation by TM.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

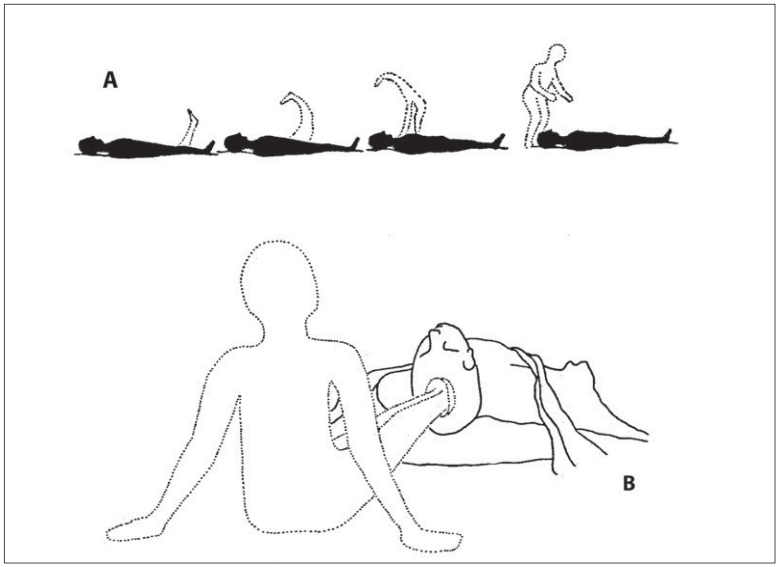

Fig. 2-6b. Kinematics of the phenomenal body-image during OBE onset. An alternative, but equally characteristic motion pattern, as described by Swiss biochemist Ernst Waelti (1983).

The classic OBE contains two self-models, one visually represented from an external perspective and one forming the center of the phenomenal world from which the first-person perspective originates. What makes the representationalist and functionalist analysis of OBEs difficult is the fact that there are many *related* phenomena, e.g., autoscopic phenomena during epileptic seizures, in which only the first criterion is fulfilled.[55] Devinsky et al.[56] have differentiated between autoscopy in the form of a complex hallucinatory perception of one's own body as being external with "the subject's consciousness […] is usually perceived within his body", and a second type, the classic OBE, which includes the feeling of

[55] For a neurological categorization see Peter Brugger, Marianne Regard, Theodor Landis: Illusory reduplication of one's own body: Phenomenology and classification of autoscopic phenomena. In: Cognitive Neuropsychiatry 2 (1997), pp. 19–38.

[56] Orrin Devinsky et al: Autoscopic phenomena with seizures. In: Archives of Neurology 46 (1989), pp. 1080–1088, here p. 1080.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

leaving one's body and viewing it from another vantage point. The incidence of autoscopic seizures is possibly higher than previously recognized, and the authors found a 6.3% incidence in their patient population.[57] Seizures involving no motor symptoms or loss of consciousness, which may not be recognized by the patient, may actually be more frequent than commonly thought.[58]

What function could this type of experience have *for* the organism as a whole? Here is a speculative proposal by Devinsky and colleagues:

> There are several possible benefits that dissociative phenomena, such as autoscopy, may confer. For example, when a prey is likely to be caught by its predator, feigning death may be of survival value. Also, accounts from survivors of near-death experiences in combat or mountaineering suggest that the mental clarity associated with dissociation may allow subjects to perform remarkable rescue manoeuvres that might not otherwise be possible. Therefore, dissociation may be a neural mechanism that allows one to remain calm in the midst of near-death trauma.[59]

It is not at all inconceivable that there are physically or emotionally stressful situations in which an information-processing system is forced to introduce a "representational division of labour" by distributing different representational functions into two or more distinct self-models (for instance in what in the past was called "multiple personality disorder").[60] The OBE may be an instance of transient functional modularization, of a "purposeful," i.e., functionally adequate, separation of levels of representational content in the PSM. For instance, if the system is cut off from somatosensory input or flooded with stressful signals and information threatening the overall integrity of the self-model as such, it may be advantageous to integrate the ongoing conscious representation of higher cognitive functions like attention, conceptual thought, and volitional selection processes into a *separate* model of the self. This may allow for a high degree of integrated processing, i.e., of "mental clarity," by functionally encapsulating and thereby *modularizing* different functions like proprioception, attention, and cognition in order to preserve at least

[57] Ibid., p. 1085.
[58] A case study of a patient who initially had OBEs for several years and later suffered from generalized epileptic seizures can be found in Patrik Vuilleumier u.a.: Héautoscopie, extase et hallucinations expérientielles d'origine épileptique. In: Revue Neurologique 153 (1997), pp. 115–119, here p. 116.
[59] Devinsky et al., 1989, p. 1088.
[60] See Metzinger, 2003a, Section 7.2.4.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

some of these functions in a life-threatening situation. Almost all necessary system-related information is still globally available, and higher order processes like attention and cognition can still operate on this information as it continues to be presented in an integrated manner, but its distribution across specific sub-regions of phenomenal space as a whole changes dramatically. Only one of the two self-models is truly "situated" in the overall scene; only one of them is immediately embodied and virtually self-present in the sense of being integrated into an internally simulated behavioral space.

It has long been known that OBEs occur not only in healthy subjects, but in certain clinical populations (e.g., epileptics) as well. In a recent study, Olaf Blanke and colleagues were able to localize the relevant brain lesion or dysfunction in the temporoparietal junction (TPJ) in five out of six patients. It was also possible, for the first time, to induce an OBE-type state by direct electrical stimulation. These researchers argue that two separate pathological conditions may be necessary to cause an OBE. First, a disintegration in the self-model or "personal space" (brought about by a failure to integrate proprioceptive, tactile, and visual information regarding one's own body) plus an additional, second disintegration between external, "extra-personal" visual space, and the internal frame of reference created by vestibular information. The experience of seeing one's own body in a position that does not coincide with its felt position could therefore be caused by cerebral dysfunction at the TPJ, causing both types of functional disintegration and thereby leading to the representational configuration described above.

Using evoked potential mapping, these authors also showed that a selective activation of the TPJ takes place 330–400ms after healthy, willing volunteers mentally imagined themselves being in a position and taking a visual perspective characteristic of an OBE. At the same time, it is possible to impair this mental transformation of the bodily self-model by interfering at this specific location with TMS. In an epileptic patient with OBEs caused by damage at the TPJ, it could be shown that by mimicking the OBE-PSM (i.e., by mentally simulating an OBE like the ones she had experienced before), there was a partial activation of the seizure focus.[61] Therefore, there exists an anatomical bridge overlap between these three very similar types of phenomenal mental content.

[61] See Olaf Blanke: Multisensory mechanisms of bodily self-consciousness. In: Nature Reviews Neuroscience 13 (2012), pp. 556–571.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

What is most needed at the current stage is an experimental design that makes OBEs a controllable and repeatable phenomenon in healthy subjects, under laboratory conditions. Achieving this interim goal would be of great interest, not only from an empirical, but also from a philosophical perspective. Studying the functional fine structure of embodiment by developing a convincing representationalist analysis of phenomenal *disembodiment* would certainly shed new light on the issue of non-conceptual self-awareness and the origin of a conscious first-person perspective. In particular, it would be of high theoretical relevance if one could empirically demonstrate the possibility of minimal selfhood *without an agency component*.[62] Let me therefore give you a brief example of my own recent research. Example no. 5 is a study based on interdisciplinary cooperation between neuroscience and philosophy of mind, and, specifically, on an experimental design originally developed from philosophical considerations.[63]

The classical Rubber-Hand Illusion (Fig. 4) only tells us something about the target property of "ownership" (of body parts), but not about "selfhood" (ownership of the *whole* body, so to speak). To manipulate attribution and localization of the *entire* body and to study selfhood *per se*, we designed an experiment based on clinical data in neurological patients with out-of-body experiences. These data suggest that the spatial unity between self and body may be disrupted leading in some cases to the striking experience that the conscious self is located in an extra-bodily position. Therefore, the aim of the present experiments was to induce out-of-body experiences in healthy participants in order to investigate the phenomenal target property of selfhood. We hypothesized that under adequate experimental conditions, participants would experience a visually presented body as if it was their own, inducing a drift of the subjectively experienced bodily self to a position outside one's bodily borders. Can one create a whole-body analog of the RHI, an illusion during which healthy participants experience a virtual body as if it were their own and locate their self outside their body's boundaries at a different position in space?

Bigna Lenggenhager and Tej Tadi applied virtual reality to examine the possible induction of out-of-body experiences by using multisensory conflict. In the first experiment, participants viewed the back of their body

[62] See Blanke, Metzinger (footnote 25).

[63] For details see Lenggenhager et al., 2007. For a popular scientific overview, see Metzinger [2014] (footnote 2). For an excellent view of the further development of this field, see Blanke [2012] (footnote 61).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

filmed from a distance of two meters and projected onto a 3D-video *head-mounted display* (HMD; see Fig. 7). The participants' back was stroked during one minute either synchronously or asynchronously with respect to the virtually seen body. Global self-attribution of the virtual character was measured by a questionnaire that was adapted from the RHI. Global self-localization was measured by passively displacing the blindfolded participants immediately after the stroking and asking them to return to their initial position (Fig. 7).

While being stroked, the subjects were either shown their own back ("own body condition"), the back of a mannequin ("fake body condition") or an object ("object condition") being stroked and projected directly (synchronously) or with a time lag (asynchronously) onto a HMD. After being stroked, the subjects were passively displaced and then asked to return to their initial position and fill out a modified "rubber-hand questionnaire".

Results of the questionnaire showed that for the synchronous "own body" and "fake body" conditions, subjects often felt as if the observed virtual figure were their own body. This impression was less likely to occur in the "object condition" and in all of the asynchronous conditions. The synchronous experimental conditions also showed a significantly larger shift towards the projected real or fake body than the asynchronous and control conditions. These data suggest that self-location — due to conflicting visual-somatosensory input — is as prone to misidentification and mislocalization as was previously reported for body parts, as in the RHI.

Illusory self-localization to a position outside one's body shows that bodily self-consciousness and selfhood can be dissociated from an accurate representation of one's physical body position. This differs from the RHI where the aspect of selfhood remained constant and only the attribution and localization of the stimulated hand was manipulated. Does illusory self-localization to a position outside one's body mean that we have experimentally induced full-blown out-of-body experiences? No, this was only a first step and the effect is also much weaker than in the Rubber-Hand illusion. For many subjects, it was much more similar to the phenomenology of a heautoscopy. But it is quite clear what the next steps will have to be. Out-of-body experiences are characterized by disembodiment of the self to an extracorporeal location, an extracorporeal visuospatial perspective (although there are also purely auditory OBEs), and the seeing of one's own body from this extracorporeal self-location. As the present illusion was neither associated with overt *disembodiment* – that is, with the feeling of having left or lost one's body – nor with a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

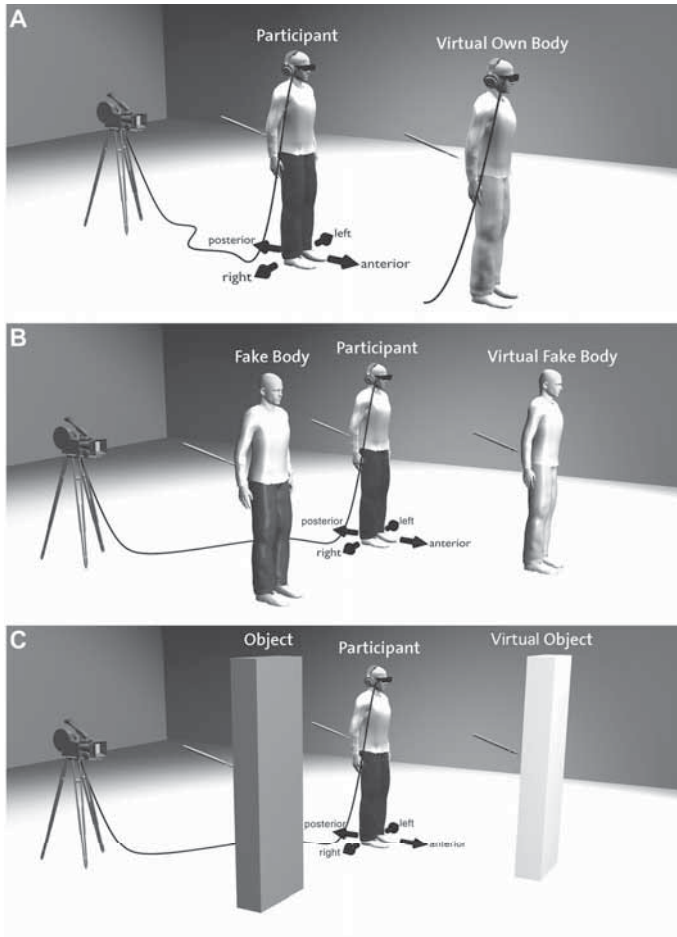

Fig. 2-7. *The creation of a full-body variant of the Rubber-Hand illusion.* (A) Participant (in dark trousers) looking through a HMD (a *head-mounted display* is a head mounted visual output device, which projects the images generated by a computer onto a nearby screen, or even directly onto the retina), sees his own virtual body (lighter trousers) in 3D, standing two meters in front of him and being stroked synchronously or asynchronously in the participant's back. In other conditions the participant sees either (B) a virtual fake body (namely the back of a mannequin, bright trousers) or (C) a virtual non-corporeal object being stroked synchronously or asynchronously in the back. Dark colors indicate the actual location of the physical body/object, whereas light colors represent the virtual body/object seen on the HMD. (Illustration by M. Boyer)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

change in visuospatial perspective, we argue that we have induced only some aspects of out-of-body experiences or rather the closely related experience of heautoscopy that has also been observed in neurological patients.[64] These are characterized, as in the case of autoscopy, by an illusory visual representation of the body – a visual *doppelgänger* – as well as a rapid change of the self-localization between the illusory and the real body.

To give just one example for further research: I believe that an additional necessary condition involved in generating full-blown out-of-body experiences and the complete transfer of selfhood to the illusory body is a transient episode of visual-vestibular disintegration. At least two spatial frames of reference must be functionally dissociated in order to have not only a "teleportation-OBE," but a realistic exit phenomenology, a gradual motion path through phenomenal space (see Fig. 6a). Setting aside cases of sudden OBEs, as after accidents or direct electrical stimulation of the brain, the exit phenomenology at the beginning of experience should reflect this decoupling of different spatial references at the level of consciousness. Why is this principle relevant from a theoretical perspective, and why is it difficult to test experimentally? In standard situations, and as opposed to all other conscious models of aspects of reality, the human PSM is anchored in the brain through a continuous flow of self-generated input.[65] There exists a persistent causal link into the physical body itself: it is the interoceptive input from intestines and blood vessels, the constant flow of information from the vestibular system and also the emotional background state, that firmly anchor the human PSM in its neural basis and make it hard, so to speak, "to copy it out" of the body or to move it to another carrier system.[66] In order to understand the SMT better, we must turn to this point now — it explains why our conscious model of reality is a *centered* model of reality.

[64] See the original publication for further references, Metzinger [2014] and Blanke [2012] (footnote 61).
[65] I have explained this point in more detail in "*Being No One*" (see T. M. [2003] [footnote 4]).
[66] See Seth (footnote 11).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Step Four: the Bodily Self as a Functional Anchor of Phenomenal Space

Above, I drew attention to the distinction between the representational and the functional analysis of the first-person perspective. The central theoretical problem on the functional level of description can be summed up by the following question: what exactly is the difference between the PSM and the other phenomenal models that are currently active in the system? Is there a characteristic causal mark of the PSM? Which *functional property* is responsible for turning it into the stable center of phenomenal representational space?

This is my first, preliminary, answer. The self-model is the only representational structure that is anchored in a *continuous source of internally generated input* in the brain. Let us call this the "persistent causal link hypothesis." Whenever conscious experience arises (i.e., whenever a stable, integrated model of reality is activated), this continuous source of internal proprioceptive input also exists.[67] The human self-model possesses an enduring causal link in the brain. It has parts, which in turn are realized by *permanent* forms of information processing on *permanent* forms of self-generated input and low-level autoregulation. To put this general point differently, the body, in certain of its aspects, is the only perceptual object from which the brain can never run away. There are a number of obvious candidates for sources of high invariance. For example, the following four different types of internally generated information that constitutes a persistent functional link between the PSM and its bodily basis in the brain during conscious episodes:

---

[67] The importance of interoception for self-consciousness is investigated in a number of recent studies by Manos Tsakiris and his colleagues; see Manos Tsakiris, Ana T. Jiménez, Marcello Costantini: Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body representations. In: Proceedings of the Royal Society B 278 (2011), pp. 2470-2476; Vivien Ainley et. al.: Looking into myself: Changes in interoceptive sensitivity during mirror self-observation. In: Psychophysiology 49 (2012), pp. 1672-1676; an important recent study comes from Jane E. Aspell et al.: Turning body and self inside out: Visualized heartbeats old bodily self-consciousness and tactile perception. In: Psychological Science 24 (2013), pp. 2445 to 2453. For an interesting computational model, that works well with a theory of self-consciousness, Anil K. Seth, Keisuke Suzuki, Hugo D. Critchley: On interoceptive predictive coding model of conscious presence. In: Frontiers in Psychology 2 (2012), Art. 395, see especially Seth (footnote 11).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

- Inputs from the vestibular organ: the sense of balance.
- Inputs from the autonomously active, invariant part of the body schema: the continuous "background feeling" in the spatial model of the body, which is independent of external input, e.g., via motion perception.
- Inputs from the visceral sensors, but also from the blood vessels, for instance from the cardiovascular mechanosensors: "gut feelings" and somato-visceral forms of self-presentation.
- Inputs from certain parts of the upper brain stem and hypothalamus: background emotions and moods, which are anchored in the continuous homeostatic self-regulation of the "internal milieu", the biochemical landscape in our blood.

Philosophically, it is not so much the neurobiological details that are crucial, but rather the highly plausible assumption that there is a certain part of the human self-model that is characterized by a high degree of stimulus correlation, and that depends exclusively on internally generated information. This layer of the PSM is directly and permanently anchored in stimuli from the inside of the body. This fact is phenomenologically relevant since it makes the decisive contribution to the quasi-Cartesian phenomenology of substantiality ("I am an ontologically autonomous entity who can hold itself in existence") and the self-knowability of the subject ("I know that I *know* that I myself exist"). But it is also epistemologically relevant, since it functionally fixes the guaranteed reference mentioned above. Do you still remember patient AZ from example no. 2? The weaker degree of phenomenological "vividness" or "realness" in her phantom limbs may reflect exactly the absence of permanent bottom-up stimulation that in normal situations is caused by existing physical limbs. In this context, Marcel Kinsbourne has spoken of a "background 'buzz' of somatosensory input" [68]. To capture the phenomenology involved in this sheer "raw feel of embodiment" on the representationalist level of description, I like to distinguish between self-presentation and self-representation. [69] Phenomenologically, the first

[68] Marcel Kinsbourne: Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In: José Luis Bermúdez u.a. (Ed.): The Body and the Self. Cambridge 1995, p. 217.

[69] For an extensive theoretical treatment of the subject and numerous recent empirical results on the body as an anchor of conscious experience, see Damásio (1999). António Damásio uses the term of a core self, and elsewhere (Metzinger,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

concept is related to the purely sensory feeling of bodily presence, which so interestingly goes along with a subjective sense of temporal immediacy and the experiential certainty of possessing direct, non-inferential self-knowledge. What exactly is this deepest layer of the phenomenal self? Why is it the origin of the first-person perspective? My hypothesis is that the constant self-organizing activity of those regions of the bodily self that are independent of external input constitutes the functional center of phenomenal representational space.

As in our first example of how to understand the concept of a self-model, we used the experiment in which Ramachandran managed to mobilize a paralyzed phantom limb. A self-*presentation* is exactly that part of the phantom limb that remains conscious independently of the occurrence of movement. If *this* part is lost, you also lose the subjective experience of bodily presence – you turn into a "disembodied being"[70]. But there may even be other, more general empirical perspectives, from which the self-model is necessarily related to the baseline of brain activity per se, as it can be observed in the resting state.[71]

1993, p. 156ff; Metzinger, 2003a, Section 5.4) I introduced the technical concept of "phenomenal self-presentation" (as opposed to self-*re*presentation). On the level of body-representation, self-presentation is what AZ lacks in her phantom limbs, whereas self-representation is what she actually has — although, as the referent of this representation never existed, this obviously is also a form of misrepresentation.

[70] Again, the corresponding phenomenological state classes exist. In Metzinger (1993) and Metzinger (1997), I discussed Oliver Sacks' example of the "disembodied lady". In this context, see also the famous case of Ian Waterman, which is discussed in Metzinger (2003a). Also interestingly, the most recent dream research shows how it is possible that the phenomenal subject may be stable, but still localized only as an unextended point in a spatial frame of reference (cf. TM: Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. In: Frontiers in Psychology 4 [2013], p. 746; Jennifer M. Windt: The immersive spatiotemporal hallucination model of dreaming. Phenomenology and the Cognitive Sciences 9 [2010], pp. 295-316; and Windt [n. 25]).

[71] See D. A. Gusnard: Being a self: considerations from functional imaging. In: Conscious. Cogn. 14 (2005), pp. 679–697.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Step Five: Autoepistemic Closure – Transparency and the Naive-Realistic Self-misunderstanding

Back on the *representational* level of analysis, the central theoretical problem is that one might easily accuse me of mislabeling the actual problem by introducing the concept of a “self-model". The central feature of self-consciousness – the fact that we always see ourselves as a self, as opposed to a model – seems not to be explained by the SMT. But for starters, a self-model is not a model of a mysterious thing that we then call the *self*. It is a continuous and self-directed process tracking global properties of the organism and their temporal development. Second, at least according to certain modal intuitions, there appears to be no necessary connection between the fundamental functional and representational properties on the one hand and the *phenomenal* target properties of “mineness”, “prereflexive/preagentive selfhood” and "perspectivalness" on the other. Some think all this could easily occur without resulting in a real phenomenal self or a subjective inner perspective. As long as the term "consciousness" remains so empty of real content, as is the case in the current debate, zombies seem to remain at least logically possible: it is conceivable that biological information-processing systems could develop and successfully employ a representational space centered by a self-model *without* also developing self-consciousness. More interestingly, *even given the phenomenal level*, i.e., even in a system that is already conscious, it is not obvious or self-evident that the specific phenomenology of *selfhood* should emerge. What would, by logical necessity, bring about an ego? A "self-model" is by no means a self, but only a representation of the system as a whole – it is no more than an integrated *system-model*. If the functional property of centeredness and the representational property of having a self-model are to lead to the phenomenal property of perspectivalness, of a consciously experienced sense of self, the conscious system-model must turn into a phenomenal self. The decisive philosophical question is this: how does the existence of a functionally centered representational space necessarily lead to the emergence of a conscious self and what we commonly call a phenomenal first-person perspective? In other words, how does the system-model turn into a *self*-model?

My answer is that a genuinely conscious self emerges at the exact moment when the system is no longer able to recognize the self-model it is currently generating *as* a model on the level of conscious experience. So how does one get from the functional property of "centeredness" and the representational property of "self-modeling" to the phenomenal target property of "pre-reflexive self-intimacy"? The solution has to do with what

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

philosophers call "phenomenal transparency"[72]. The conscious representational states generated by the system are *transparent*, i.e., they no longer represent the very fact that they are models on the level of their content. Consequently – and this is only a visual-phenomenological metaphor first introduced by the British philosopher G. E. Moore – the system simply looks right "through" its very own representational structures, as if it were in direct and immediate contact with their content. Please note how this is only a statement about the system's *phenomenology*. It is not a statement about epistemology, about the possession of knowledge: you can be completely deluded and have no or very little knowledge about reality (or your own mind) and at the same time enjoy the phenomenology of certainty, of *knowing that you know*. Phenomenal transparency is not *epistemic* transparency, or Descartes' classical – and now empirically falsified – idea that we cannot be wrong about the contents of our own mind. Transparency, as defined in this context, is exclusively a property of *conscious* states. Unconscious states are neither transparent nor opaque. Phenomenal transparency is also not directly related to the third technical and philosophical term, "referential transparency." Nonlinguistic creatures, incapable of conceptual thought, can have phenomenally transparent states as well as systems that are subject to fundamental self-deception. There is no naïve realistic assumption here, no specific semantic context, neither a belief nor an intellectual attitude, but a feature of phenomenal experience itself.

I have two causal hypotheses about the micro-functional underpinnings and the evolutionary history of transparent phenomenal states. First, in a very small time-window, the neural data structures in question are activated so quickly and reliably that the system is no longer able to recognize them as such, for instance due to the comparatively slow temporal resolution of *meta-representational* functions or because they

[72] For a short explanation of the concept of "phenomenal transparency", see Thomas Metzinger: Phenomenal transparency and cognitive self-reference. In: Phenomenology and the Cognitive Sciences 2 (2003), pp. 353–393. Metzinger, 2003b is the German precursor. On the relationship between phenomenal transparency and epistemic justification, see Thomas Metzinger (2014c): How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. Cognitive Neuroscience, DOI: 10.1080/17588928.2014.905519 und Thomas Metzinger und Jennifer Windt: Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath & J. Kipper (Eds.), Experimentelle Philosophie. Frankfurt am Main: Suhrkamp.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

belong to a stable, Bayesian Optimal model, where the probability of error is already greatly minimized.[73] Introspectively, the construction process is invisible. Second, in a much larger explanatory time-window, there was apparently no evolutionary pressure on the respective parts of our functional architecture in the process of natural selection. For biological systems like us, naive realism was a functionally adequate background assumption. We needed to know "careful, there is a wolf nearby!", but not "a wolf-representation is active in my brain right now".

Phenomenal transparency has therefore still something to do with knowledge (and not only with the subjective experience *per se*): it is a special form of darkness. It is a lack of knowledge. Epistemologically speaking, it is an implicit lack of self-knowledge, of knowledge about the functional deep structure of your own mind, which is itself again not explicitly represented. As Franz Brentano[74] and more recently Daniel Dennett[75] pointed out, the representation of absence is not the same thing as the absence of representation. In transparent states, there is no representation of earlier processing stages or of dynamically unstable states. In the phenomenology of visual awareness, it means not being able to see something. Phenomenal transparency *in general*, however, means that the representational character of the contents of conscious experience itself is not accessible to subjective experience: we can no longer see that what we are concerned with here is a *prediction*, with a very good internal model, whose function is to avoid future surprises. This analysis can be applied to all of the sensory modalities, especially to the integrated phenomenal model of the world as a whole. Because the very *means* of representation cannot be represented as such, the experiencing system necessarily becomes entangled in naive realism; it experiences itself as being directly in contact with the contents of its own conscious experience. It is unable to experience the fact that all of its experiences take place in a *medium* – and this is exactly what we mean by the "immediacy" of phenomenal consciousness. In a completely transparent representation, the very mechanisms that lead to its activation as well as the fact that its contents depend on a concrete inner state as a carrier can no longer be recognized by way of introspection. As analytic philosophers classically like to say: "only content properties are introspectively accessible, vehicle

[73] See Howhy 2013 and Friston 2010.

[74] See Brentano, F. (1973) [1874]. Psychologie vom empirischen Standpunkt. Erster Band. Meiner, Hamburg, p. 165f.

[75] Daniel Dennett: Consciousness Explained. Boston 1991, p. 359.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

properties (i.e. the properties of the representational carrier) are inaccessible". Therefore, the phenomenology of transparency is the phenomenology of naive realism.

Many phenomenal representations are transparent because their content and its very existence appear to be fixed in all possible contexts. According to subjective experience, the book you are currently holding in your hands will always stay the same book – no matter how the external perceptual conditions vary. You never have the experience that an "active object emulator" in your brain is currently being integrated into your global reality-model. You simply experience the *content* of the underlying representational process: the *book* as effortlessly given, here and now. The best way to understand the concept of transparency is to distinguish between the vehicle and the content of a representation, between representational carrier and representational content.[76]

The representational carrier of your conscious experience is a particular brain process. This process – that itself is in no way "book-like" – is not consciously experienced; it is transparent in the sense that, phenomenologically, you look right through it. What you look at is its representational content, the perceptually mediated existence of a book, here and now. When looked at from the conceptual external-perspective, this content is an abstract property of a concrete representational state in your brain. If the representational carrier is a good and reliable instrument for the generation of knowledge, its transparency allows you to "look right through" it out into the world, at the book in your hands. It makes the information it carries globally available without you – the person as a whole – having to worry about *how* this actually happens. What is special about most phenomenal representations is that you experience their content as maximally *concrete* and unequivocal, as directly and immediately given even when the object in question – the book in your hands – does not really exist at all, but is only a hallucination. Phenomenal representations appear to be exactly that set of representations for which we cannot distinguish between representational content and representational carrier on the level of subjective experience.

There are counterexamples, of course, and they may help to illustrate further the concept of "transparency." For instance, *opaque* phenomenal representations arise when the information that their content is the result of an internal representational process suddenly becomes globally available. If you suddenly discover that the book in your hands does not really exist,

[76] See also Dretske 1998, pp. 45ff.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the hallucination turns into a pseudo-hallucination. The information that you are not looking at the world, but rather "at" an active representational state that apparently is not functioning as a reliable instrument for the generation of knowledge at this moment now also becomes available, and it does so on the level of subjective experience itself. The phenomenal book state becomes opaque. You lose *sensory* transparency. You become aware of the fact that your perceptions are generated by your sensory system and that this system is not always completely reliable. Not only do you now suddenly experience the book as a representation, you also experience it as a *mis*representation.

Let us further assume that you suddenly discover that not only your perception of the book, but all of your philosophical thoughts about the problem of consciousness are taking place in a dream. Then, this dream would turn into a lucid dream.[77] The fact that you are currently not experiencing a world, but only a *world-model* would become globally available; now, you could use this information to control your actions, thoughts, and the direction of attention. You would lose *global* transparency. The interesting point, however, is that cognitive availability alone is not sufficient to dissolve the naive realism of phenomenal experience. You cannot simply "think" yourself out of your phenomenal model of reality by changing your opinions about this model: the transparency of phenomenal representations is cognitively impenetrable; here, phenomenal knowledge is not the same as conceptual/propositional knowledge.

Now, the final step is to apply this insight to the self-model. Here is my key claim – we are systems that are experientially unable to recognize our own sub-symbolic self-model *as* a self-model. For this reason, phenomenologically, we operate under the conditions of a "naive-realistic self-misunderstanding": we experience ourselves as being in direct and immediate epistemic contact with ourselves. By logical necessity – this is

[77] For a discussion of the reasons for regarding lucid dreams as a philosophically relevant class of conscious states, see Metzinger, 2003a, Section 7.2.5; more on the topic can be found in Windt and Metzinger, 2007 The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In: Deirdre Barrett, Patrick McNamara (Hg.): The New Science of Dreaming. Westport, CT 2007; Valdas Noreika u.a.: New perspectives for the study of lucid dreaming: From brain stimulation to philosophical theories of selfconsciousness. Commentary on „The neurobiology of consciousness: Lucid dreaming wakes up“ by J. Allan Hobson. In: International Journal of Dream Research 3 (2010), S. 36–45; and especially in Windt 2015.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the conceptual point –, a phenomenally transparent self-model will create the experience of *being infinitely close to yourself*. The core of the self-model theory is that this is how the basic sense of selfhood arises and how a phenomenal self that is untranscendable for the respective system comes about.[78] The content of non-conceptual self-consciousness is the content of a transparent PSM. It also commits me to a specific prediction: were the PSM to lose its transparency and become opaque, were the organism as a whole capable of recognizing its current self-model *as* a model, then the phenomenal property of selfhood would disappear. In standard phenomenological configurations, however, the entity that looks at the book in its hands is itself a form of transparent phenomenal content. And this is also true of the "at"-ness inherent in this act of visual attention, of the relation that seems to connect subject and object.

Step Six: the PMIR — the Phenomenal Model of the Intentionality Relation

Let us take one more step before we close. The experience of selfhood is intimately related not only to the sense of ownership, but also to the experience of agency. It is not only a question of having a transparent self-model, but also of directedness, of being dynamically related to target objects and goal states. Here are two further examples, this time from yet another academic discipline — experimental neuroscience using macaque monkeys as subjects.

Classical neurology hypothesized about a "body schema", an unconscious but constantly updated map of body shape and posture in the brain.[79] Recent research shows how Japanese macaque monkeys can be

[78] To identify, on the conceptual and empirical level, the minimal constitutive conditions for the emergence of phenomenal self (relative to a specific class of systems) is one of the main research goals of the self-model theory of subjectivity. This research program is tied to the concept of *minimal phenomenal selfhood* (MPS; Blanke, Metzinger [footnote 25]). A good candidate is the phenomenally transparent self-localization in a spatial and temporal reference, for which agency and the presence of an explicit body model seem not to be necessary conditions (see Metzinger [2013] (n. 70) and Windt [n. 25, 71]).

[79] The terminology was never entirely clear, but it frequently differentiated between an unconscious "body schema" and a conscious "body image". For a philosophical perspective on the conceptual confusion surrounding both notions, see Gallagher (2005); for an excellent review of the empirical literature, see Maravita (2006). For An excellent recent overview see Atsushi Iriki , Osamu

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

trained to use tools even though they only rarely exhibit tool-use in their natural environment.[80] During successful tool-use, changes take place in specific neural networks in their brains – a finding that suggests that the tools are temporarily integrated into their body schema. When a food pellet is dispensed beyond the reach of their hands and they skillfully use a rake to pull it closer, one can observe a change in their bodily self-model in the brain. In fact, it looks as if their conscious model of their hand has been expanded towards the tip of the tool. A more precise way of describing what happens is to say that, on the level of the monkey's conscious model of reality, properties of the hand are now transferred to the distant tip of the tool. We are acquainted with the same effect in human beings. In our own case, repeated practice can turn the tip of a tool into a part of our own hand, a part that can be used just as "sensitively" and as skillfully as our own fingers.

In other words, recent neuroscientific data clearly support the view that tools do not just enable us to extend our spatial reach. They show that any successful extension of behavioral space is also mirrored in the neural substrate of the body image in the brain. The brain constructs an "internalized" image of the tool by swiftly assimilating it into the existing body image as a whole. We do not know, of course, whether monkeys actually have the conscious experience of ownership or only the unconscious mechanism. But we do know about several similarities between macaques and humans that make this assumption seem plausible. This may be the very beginning of mentally *simulating* yourself as currently being directed at a target object or goal state. And this leads us to second major aspect of selfhood: besides global *ownership* what we need to understand is *agency* – global control.

One exciting aspect of these new data is that they shed light on the evolution of tool-use. A necessary precondition of expanding your space of action and your capabilities by using tools clearly seems to be the ability to integrate them into a preexisting self-model. You can only engage in goal-directed and intelligent tool-use if your brain temporarily represents them as part of your own self. Intelligent tool-use was a major achievement in human evolution. One may plausibly assume that some elementary building block of human tool-use abilities already existed in

---

Sakura: The neuroscience of primate intellectual evolution: Natural selection and passive and intentional niche construction . In: Philosophical Transactions of the Royal Society B 363 (2008) , pp 2229-2241.

[80] See Angelo Maravita, Atsushi Iriki: Tools for the body (schema). In: Trends in Cognitive Sciences 8 (2004), pp. 79–86.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the brains of our ancestors. Then, due to some not-yet-understood evolutionary pressure, it rapidly expanded into what we see in humans today.[81] I think one of the most interesting aspects of recent socio-cultural development consists precisely in the fact that people are attempting to instrumentalize *one another* to an hitherto unknown extent with the help of their new self-models: namely, people try very often to enlarge their own sphere of action by controlling *other* people and using them as tools. My speculative thesis is that this has driven the evolution of large companies in a decisive way.

There is a new, rapidly growing field of research in which engineers and neuroscientists work together: *brain–machine interfaces*[82]. One application of this general idea consists in driving and controlling artificial limbs or robotic manipulators with the help of ensembles of cortical neurons, allowing a machine to carry out motor commands generated in the brain. Meanwhile, monkeys are already able to remotely control the steps of humanoid robots on the other side of the world (from *Duke University* in the United States to the *Computational Brain Project* at the *Japan Science and Technology Agency* in Japan) via the Internet and in real time, exclusively through the recording of their brain activity. Miguel Nicolelis writes:

> The most stunning finding is that when we stopped the treadmill and the monkey ceased to move its legs, it was able to sustain the locomotion of the robot for a few minutes – just by thinking – using only the visual feedback of the robot in Japan.

In our context, perhaps the most interesting observation in this experiment[83] is how the monkey gradually begins to neglect his original arm, which is, after all, a part of his biological body. That is, as he now tries to control the feedback in a new kind of motor task and with a different goal-state, optimizing a new set of motor parameters by trying to

---

[81] See Atsushi Iriki, Michio Tanaka, Yoshiaki Iwamura: Coding of modified body schema during tool-use by macaque post-central neurons. In: Neuroreport 7 (1996), pp. 2325–2330; and Maravita, Iriki, 2004.

[82] For a brief overview, see Mikhail Lebedev, Miguel Nicolelis: Brainmachine interfaces: past, present, and future. In: Trends Neurosci 29 (2006). H. 9, pp. 536–546.

[83] For details, see Jose M. Carmena u.a.: Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates. In: PLoS Biology 1 (2003), pp. 193– 208.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

control a real-world robot arm or even a virtual arm he sees on the screen in front of him, his brain seems to undergo certain changes. He optimizes a new set of motor parameters by attempting to control a real robot or even a virtual arm he sees on the screen in front of him. The "tuning properties" of neurons change. Here is how Lebedev and Nicolelis[84] describe the effect:

> Remarkably, after these animals started to control the actuator directly using their neuronal activity, their limbs stopped moving, while the animals continued to control the actuator by generating proper modulations of their cortical neurons. The most parsimonious interpretation of this finding is that the brain was capable of undergoing a gradual assimilation of the actuator within the same maps that represented the body.

From the perspective of SMT, the self-model theory, the most plausible interpretation is that, once the monkey has successfully embedded an internal representation of this new actuator into his conscious self-model, the representations of his old body parts lose certain functional properties. They transiently become less and less available for attentional processing and gradually recede from conscious experience. These examples teach us two further important insights. Self-evidently, the PSM is an important part of a *control hierarchy*; it is a means to monitor certain critical aspects of the process by which the organism generates flexible, adaptive patterns of behavior; second, it is highly plastic in the sense that multiple representations of objects *outside* the body can transiently be integrated into it. This is not only true of rubber-hands, but even more so of tools in the most general sense – extensions of bodily organs which must be successfully controlled in order to generate intelligent, goal-directed behavior. The self-model is the functional window through which the brain can interact with the body *as a whole*, and vice versa. If the body is augmented by sticks, stones, rakes, or robot arms, the self-model itself has to be extended. If an integrated representation of body-plus-tool is in existence, the extended system of body-plus-tool can become part of the brain's control hierarchy. Put in another way: how could one learn to use a tool intelligently – that means: in a flexible and context-sensitive manner – *without* after all integrating it into the conscious self? The conscious self-model is a virtual organ that allows us to *own* feedback loops, to initiate, sustain, and flexibly adapt control processes. Some elements of the control loop are physical (such as

[84] Lebedev, Nicolelis, p. 542.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the brain and tools); others are virtual (such as the self-model and goal-state simulation).

Under SMT, to embed complete robotic or virtual body (avatars) into the PSM and thereby to exercise causal control is also clearly conceivable. An empirical prediction among the conceptual assumptions of SMT is that it must be possible in principle to link the human self-model, bypassing the non-neural, biological body in a causally more direct way to artificial effectors and sensors and thereby functionally situate it in a technologically produced environment. In an ambitious pilot study, Ori Cohen, Doron Friedman and their colleagues have demonstrated that it is possible with functional real-time magnetic resonance imaging to read out the motor intentions of a subject and to transmit them directly as high-level motor commands to a humanoid robot, which are then transformed into physical actions. Meanwhile the subject can experience the entire experiment visually through the eyes of the robot.[85] The process is based on pictorial representations of movements that allow the subjects "to act directly with its PSM"[86] by controlling a humanoid robot HOAP3 in France from a scanner in Israel.

I have already pointed out above that people (and also some other animals) often want to control the behavior or condition of *another* person. We "instrumentalize" and "seize" each other, sometimes we even make "serfs." Human beings are constantly enlarging themselves – not only with sticks, stones, rakes or robotic arms, but also with the brains and bodies of *other* people.[87]

---

[85] Ori Cohen et al.: MRI-based robotic embodiment: A pilot study. IEEE International Conference on Biomedical Robotics and Biomechatronics. Rome 24–27 June 2012.

[86] Elsewhere I have examined the ethical implications of this historical novel forms of action and introduced the concept of a "PSM-action", an act in which a human actor uses parts of his self-model in offline simulation, which then bypassing the proximal, non-neuronal embodiments leads to consequences in the world outside the biological body. It is conceivable, for various reasons, that the autonomy – and therefore also the ethical responsibility – of the actors, in the sense of impulse control and termination options, is limited in such situations (see TM: Two Principles for Robot Ethics in E. Hilgendorf, J.P. Günther [Ed.]. Robotik und Gesetzgebung. Baden-Baden, 2013, pp. 263-302).

[87] Metzinger, Gallese (footnote 4).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

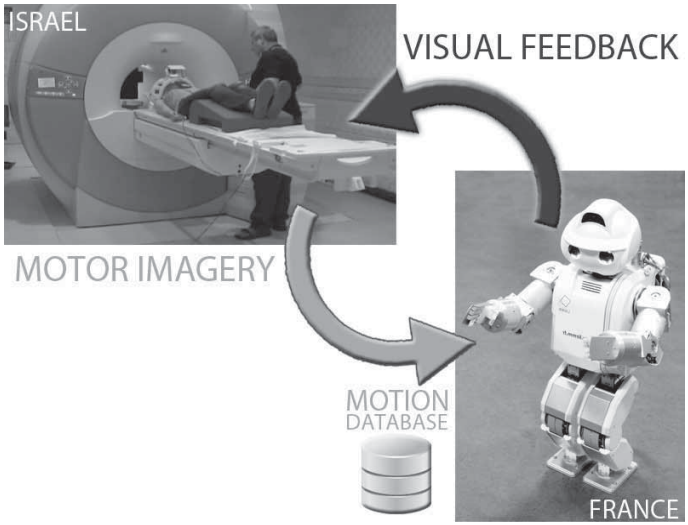

Fig. 2-8. Direct action with the PSM through *robotic re-embodiment*: the aim of the experiment was to enable a subject in Israel to control a robot in France over the internet through "direct mind control". A video demonstration can be found at http://www.youtube.com/user/TheAVL2011. (image courtesy of Doron Friedman). See further in Metzinger (2014) (footnote 2), pp. 339ff.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

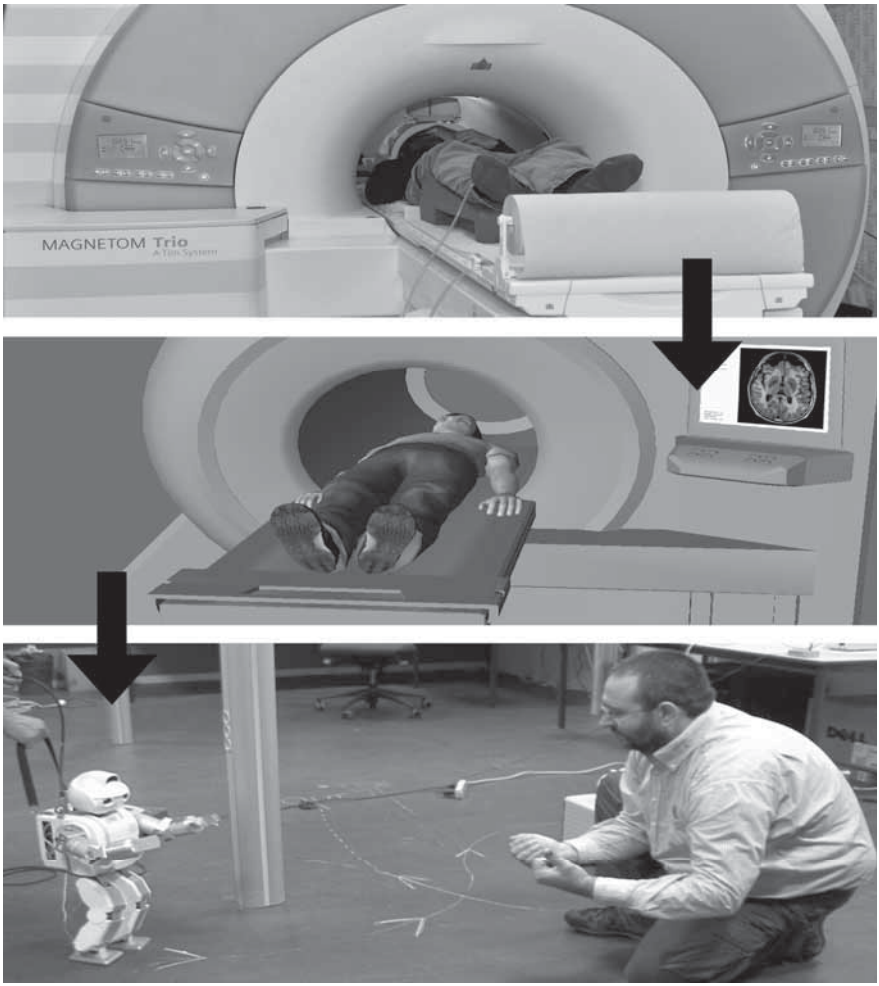

Fig. 2-9. One subject is in a NMRI (nuclear magnetic resonance imaging) scanner at the Weizmann Institute in Israel. Using data glasses, he sees an avatar who is also in the scanner. The aim is to create the illusion in the subject that he is embodied in this avatar. The motor intentions of the subject are then translated into commands that enable the avatar to move. After a training stage, the subjects at that location were able to control "directly with their minds" a distant robot in France via the Internet, where they could see the environment in France via the robot's eye camera. (image courtesy of Doron Friedman). See also Cohen 2012.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The transition from biological to cultural evolution is, without question, directly linked to the implementation of new and specific functional properties through the PSM of primates. This is one of the most interesting questions for future research: what exactly was the change in the PSM of *Homo sapiens*, as opposed to the PSM of the chimpanzee, which led to the explosion of culture and the development of complex societies? My second speculative working hypothesis would here be that it was not only the use of complex tools in the social dimension *per se*, but also the ability to have a much larger part of the control hierarchy run *offline,* and to use it in simulations while simultaneously generating an opaque (i.e., non-transparent) PSM.[88] It was the ability to represent themselves consciously *as* representing, to represent themselves *as* directed to a target state. It was the difference between the bare owning of a first-person perspective and the mental ability to represent this explicitly.

Now let us take a look at the representational architecture underlying the subjective experience of directedness in general. Phenomenologically, a transparent world-model gives rise to a reality. A transparent system-model gives rise to a self that is embedded in this reality. If there is also a transparent model of the transient and constantly changing relations between the perceiving and acting self and the objects and persons in this reality, this results in what I called a "phenomenal first-person perspective" above. A genuine inner perspective arises if and only if the system represents itself *to itself* as currently interacting with the world, and if it does not recognize this representation as a representation. It has now a conscious model of the intentionality relation (a PMIR).[89] It represents itself as directed towards certain aspects of the world. Its phenomenal

[88] The SMT says that the content of a transparent state is experienced as a given, whereas an opaque one is self-constructed (see Metzinger [2014] [footnote 2]). An interesting observation made by Jennifer Windt is that the early tools of our ancestors were not only used, but – before embedding into the self-model – they had to first be made. Monkeys, on the other hand, never or only rarely make their own tools. Our ancestors' first tools were apparently made in a few steps (four to five stone chips); complex tools, which could then be used for various activities and also kept for a longer period of time (a prerequisite for their geographical spread) required more thought out steps (and various other auxiliary tools) – and so probably also mental *offline*-simulations. *Trial and error* would hardly have led to this complexity and also to the efficiency of the production. Rather, something was actually embedded in the self-model which had been previously constructed by our own physical action and a concrete and perceptible *self*.

[89] See Metzinger (2006), especially the second part.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

space is a *perspectival* space, and its experiences are *subjective* experiences.

The intentionality relation is primarily an epistemic relation between subject and object: a mental state becomes a carrier of knowledge in virtue of being directed at something other than itself – like an arrow pointing from a person's mind to an object in the real, or even just in a possible, world. Philosophers say that this type of mental state has *intentional content*. Its content is what the arrow is pointing at. This may be an image, a proposition or even the goal of an action — as philosophers say, there is "practical intentionality" in terms of you being directed at certain "satisfaction conditions" (e.g., an action goal), and there is "theoretical intentionality" in terms of being directed at the "truth conditions" (e.g., of a sentence, i.e., an epistemic goal). If many of these arrows are consciously available, represented by the brain on the functional level of global availability, this results in a temporally extended first-person perspective. In short, it is one thing to be a biological organism that represents the world, and it is another thing to consciously represent yourself as *representing*, in "real-time" and while this is actually happening. SMT wants to understand the latter case. Now, there is not only a neurobiologically anchored core self, a self-presentation, but also a dynamic phenomenal simulation of *the self as subject* embedded in the world via constantly changing epistemic relations and agentive interactions. There is much more to be said about the central notion of a PMIR, of course.[90] But the core idea is as follows: a conscious human being is a system that is capable of dynamically *co-representing* the representational relation while representational acts are taking place, and the instrument it uses for this purpose is the PMIR. The phenomenal model of the intentionality relation (PMIR), is just another naturally evolved virtual organ, just like the PSM. The content of higher order forms of self-consciousness is always relational: the self *in the act of knowing*[91], the *currently acting* self. The ability to co-represent this intentional relationship itself while actively constructing it in interacting with a world is what it means to be a subject.

---

[90] Of course, the theory of the PMIR is more complex than I can explain in this brief overview. Apart from Metzinger (2003a), I recommend Section 4 of Metzinger (2005a, p. 26ff) for readers interested in the idea. A more detailed discussion, specifically applied to the representational architecture of conscious volitional acts, can be found in Metzinger (2006a).

[91] See Damásio, 1999, p. 168ff.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The way we subjectively experience this subject–object relation is admittedly a simplified version of the actual processes – in a sense, it is a functionally adequate confabulation. Once again, evolution favored a simple but elegant solution. The virtual self moving through the phenomenal world does not have a brain, a motor system, or sensory organs: certain parts of the environment appear directly in its mind; the perceptual process is experienced as effortless and immediate. Body movements also appear to be caused "directly." Such effects are typical for *our* type of subjective experience and – seen as a neurocomputational strategy – they have the advantage of creating a user-friendly interface. What was defined as "transparency" above is a way of describing the *closed structure* of this multimodal, high-dimensional user interface – the brain's user surface. The phenomenal self is the part of this interface that the system uses to experience *itself* as a whole, to represent itself as a thinking, knowing self and as an agent. This virtual agent "sees with his eyes" and "acts with his hands." He does not know that he has a visual or a motor cortex. The PSM is the interface that the system uses to functionally appropriate its own hardware, to control *its own* low-level dynamics and to become *autonomous*. The intentional arrows connecting this agent to objects and other selves in the currently active reality-model are phenomenal representations of transient subject–object relations – and frequently they too cannot be recognized as representational processes. In standard situations, the consciously experienced first-person perspective is the content of a transparent PMIR.

All this takes place within a phenomenal window of presence. The contents of phenomenal experience not only create a world, they also create a *present*.[92] In a sense, the core of phenomenal consciousness is just the creation of an island of presence within the physical flow of time.[93] Minimal self-consciousness is self-localization within a spatial and a temporal frame of reference[94]: Experiencing means "being here" (immersion in a scene), and this necessarily includes "being now" (presence). To be self-conscious is to be present. It means processing information in a very specific way. It means repeatedly and continuously binding discrete events that have already been represented as such into temporal gestalts, into a consciously experienced moment. Many recent

[92] See Metzinger, 2003a, Section 3.2.2.
[93] See Eva Ruhnau: Time-Gestalt and the observer. In: Thomas Metzinger (Ed.): Conscious Experience. Paderborn 1995; and the references given there, especially to the work of Ernst Pöppel.
[94] See Blanke, Metzinger (footnote 24).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

empirical data clearly demonstrate that in a certain sense, the consciously experienced present is a *remembered* present. In this sense, even the phenomenal "now" is a representational construct, a *virtual* present. And this finally helps understand what it means to say that phenomenal space is a virtual space: its content is a *possible* reality, a prediction.[95] The realism of phenomenal experience arises because it represents a possibility – the best hypothesis there is at a given moment – as an untranscendable reality or an *actuality*. In other words, the mechanisms creating temporal experience and our subjective sense of presence are transparent as well. Then, finally, this point also has to be applied to the special case of self-modeling because the virtual character of *both* the self-model *and* the window of presence are not available on the level of subjective experience itself. The system they represent turns into a *currently present subject*.

SMT solves the homunculus problem, because we can now see how no "little man in the head" is needed to interpret and "read out" the content of mental representations. It is also maximally parsimonious, as it allows us to account for the emergence of self-consciousness without assuming the existence of a substantial self, and with the PSM simultaneously providing an alternative concept for future research.[96] Does all this mean that the self – understood ontologically – is only an illusion? On second glance, the popular concept of the "self-illusion" and the metaphor of "mistaking oneself for one's inner picture of oneself" contain a logical error: *whose* illusion could this be? Speaking of illusions presupposes someone *having* them. But something that is not an epistemic subject in a strong sense of conceptual/propositional knowledge is simply *unable* to confuse itself with anything else. Truth and falsity, reality and illusion do not exist for biological information-processing systems at the developmental stage in question. So far, we only have a theory of the phenomenology of selfhood, not a theory of self-knowledge. I have only very briefly sketched here how a *phenomenal* first-person perspective can be the product of natural

[95] My own ideas on this point are very similar to those discussed by Antti Revonsuo: *Virtual reality* is simply the best technological metaphor for phenomenal consciousness we currently have. See Antti Revonsuo: Consciousness, dreams, and virtual realities. In: Philosophical Psychology 8 (1995), pp. 35–58; Antti Revonsuo: Prospects for a scientific research program on consciousness. In: Thomas Metzinger (Ed.): Neural Correlates of Consciousness: Empirical and Conceptual Questions. Cambridge, MA 2000; and especially Antti Revonsuo: Inner Presence. Cambridge, MA 2006.

[96] For the various ways in which one may be an anti-realist regarding "the" self, see Metzinger (footnote 3).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

evolution. Subjectivity in an *epistemic* sense, an epistemic first-person perspective is yet another step. Of course, the phenomenology of selfhood, of non-conceptual self-consciousness, is the most important precondition for this step, because it is the precondition for genuinely *reflexive*, conceptual self-consciousness. In a way, this is the whole point behind the theory: if we want to take high-level forms of subjectivity and intersubjectivity seriously, we must be modest and careful at the beginning, focusing on their origins on the level of non-conceptual content and self-organizing neural dynamics. And readers will not be surprised that the author of this chapter holds that subjective, first-person *knowledge* is precisely knowledge associated with a specific inner mode of presentation, namely as knowledge under a PMIR. What we today call the "first-person perspective" is a highly specific, neurally realized manner of presentation, an abstract data format, which has gradually arisen in the course of biological evolution. Subjectivity in the epistemological sense can at least in principle be naturalized as well – but only if we can tell a convincing evolutionary and neuroscientific story about how this representational architecture, this highly specific, indexical inner mode of presentation, could actually have developed in a self-organizing physical universe in the first place. What matters is that a naturally evolved system begins to model itself internally as an epistemic agent, and thereby simultaneously generating a new modality of knowledge with the help of the PSM and the PMIR, a new form of representation of knowledge about the world. Ultimately, and obviously, every single instance of the PSM/PMIR is identical with a specific time-slice in the continuous, dynamical self-organization of coherent activity taking place in an individual biological brain. In this ongoing process on the sub-personal level there is no agent – no evil demon that could count as the *creator* of an illusion. And there is also no enduring, sub-personal entity that could count as the *subject* of the illusion either. There is nobody *in* the system who could be mistaken or confused about anything – the homunculus does not exist. Self-consciousness is a property of the system as a whole.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER THREE

SELF-CONSCIOUSNESS AND FIRST-PERSON PERSPECTIVE

LUCA FORGIONE

Preliminaries

In general terms the philosophical area of self-knowledge is concerned with the knowledge of one's own mental states, e.g., the knowledge of one's current experiences, thoughts, beliefs, or desires. A classic problem, for instance, involves the possibility to determine what a subject is feeling or thinking at a given moment: although statements such as "I feel a tickle" or "I'm thinking about summertime" can sometimes express knowledge, among scholars there is significant disagreement over the nature of this knowledge.

As a rule, a subject can know what she is thinking or feeling, what she believes in or desires: if asked, under normal circumstances she can form a description of her own mental dimension and reach knowledge of her mental states. For the very nature of the subjective mental dimension, each subject is obviously in a better position than anybody else to identify her own mental states. Subjects seem to be *authoritative* about what they are thinking or feeling, since the method they employ is different from that available to others; in other words, subjects seem to be provided with a *first-person authority* about their own mental states.

The problem of the knowledge of one's mental states revolves around the involvement of the self-conscious subjective dimension. The fact that a subject acquires knowledge of her belief that Naples is a lovely city implies that the state is registered as her own; this is related to the question of self-consciousness proper, one of the major topics in the philosophical

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

arena[1]. One can be self-aware in several ways, each of which corresponding to one of the several senses of the term "self" (i.e., the embodied self, the ecological self, the narrative self, etc.); nevertheless, at first glance the self-consciousness at issue lies in the consciousness of ourselves and our personal mental dimension, of thoughts taking place, and feelings being experienced.

More specifically, the notion of self-consciousness that will be considered here can be referred to as *basic self-consciousness*. This possesses two specific correlated features, analysed in part 1, which are not owned by the consciousness of things other than oneself: the first regards the fact that self-consciousness is grounded on a first-person perspective, whereas the second concerns the fact that it should be considered a consciousness of the self as subject rather than a consciousness of the self as object. Both peculiarities are grounded on the possibility to use the term or concept *I*, as will be seen in parts 2 and 3, where a few epistemic and semantic features of the term or concept *I* will be analysed: the essential indexicality and the immunity to error through misidentification. The first regards the meaning of the term/concept *I* since any expression of self-awareness is based on indexical terms such as "I" or "me". The second concerns the fact that some singular judgments involving the self-ascription of mental properties are immune to error through misidentification relative to the first-person pronoun.

1. First-person Perspective and the Consciousness of the Self as Subject

The first-person perspective point can be explained by Baker's approach (2000, 2013). Two types of first-person perspectives can be distinguished: a *rudimentary first-person perspective*, manifested by many mammals and human infants, and a *robust first-person perspective*, manifested by language users who master first-personal language. The

[1] Gertler (2013, 19) stresses out that these issues are not specifically addressed by any theory of self-knowledge; nevertheless, since a theory of self-knowledge aims at explaining how the subject individuates her sensations, thoughts, or attitudes, the problems of self-awareness are certainly complementary to the philosophical questions pointed out by the self-knowledge area, if nothing else, for their reference to the subject. Since expressions of self-knowledge employ terms such as "I", as in "I feel a tickle", the problem of self-consciousness also concerns the ways in which the determination of *I*'s reference and the identification of those mental states as one's own may be achieved.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

latter is the conceptual capacity not only to recognize oneself as distinct from things other than oneself, but also to conceive oneself as oneself. A robust first-person perspective is exactly what marks the difference between a creature with a rudimentary first-person perspective, who can only be conscious of the environment, and a fully self-conscious subject. As a matter of fact, a mature human subject with a robust first-person perspective can attribute a first-person reference to herself on the basis of a self-concept, i.e., not only can she refer to herself in the first-person, but she can also attribute first-person reference to herself.

A crucial distinction is made between making first-person reference (as when Pasquale says, "I am tall") and attributing first-person reference (as when Mario says, "Pasquale wishes that he himself were tall"). With the latter case, Mario attributes to Pasquale a wish he would express by a first-person reference. The attribution of a first-person reference occurs in indirect discourse, in a "that-"clause following a psychological verb. The point is that a subject does not attribute first-person reference only to others but also to his own self, as when Pasquale says, "I wish that I were tall". A subject thinking "I am tall" can distinguish herself from others; a subject thinking "I wish that I were tall" can conceptualize that distinction: she can think of herself as herself. This ability to attribute first-person reference to oneself is the manifestation of strong first-person phenomena.

Following Baker (2013) and Matthews (1992), "I*" pronouns[2] are used reflexively to pick out the subject from her own point of view: given the close relation between the linguistic and the mental dimensions, I*-sentences are sentences containing "I*", whereas I*-thoughts are thoughts expressible by I*-sentences. By an I*-thought a subject conceives herself as herself*, and needs no third person referential device, such as a name, description or demonstrative to identify herself. As we will see in parts 2 and 3, certain semantic and epistemic features of the term/concept *I*[3] can be identified in this subject's capacity of self-identification: essential indexicality and immunity to error through misidentification. The former

[2] Castañeda (1966; 1967) employs an asterisk, or star, next to a pronoun ("he*") to attribute first-person reference to someone else, as in "Pasquale believes that he* is tall. This sentence is not true unless Pasquale expresses his belief in the first person: "I am tall". Matthews (1992) introduces the "I*" for sentences with first-person subjects in order to analyse the phenomena expressed by "I think that I* am F".

[3] Given the topic of this chapter, our interest regards the concept *I* and the relative *I-thoughts*; when no other specification is present, the *I* is to be considered in its conceptual nature.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

is relative to the meaning of the term/concept *I,* any expression of self-awareness being based on indexical terms such as "I" or "me"; the latter, on the other hand, refers to the fact that some singular judgments involving the self-ascription of mental (and physical, as will be seen) properties are immune to error through misidentification relative to the first-person pronoun (IEM). The subject formulating such judgments in given epistemic contexts cannot be mistaken as to whether it is he himself who is attributing a particular mental property to his own self.

At the same time the basic self-consciousness at issue here is also to be regarded as the consciousness of self-as-subject, or *subject self-awareness,* rather than the consciousness of the self as object. Following Kriegel (2003, 2007), it is possible to make a distinction between the consciousness of oneself *qua* object and the consciousness of oneself *qua* subject. For instance, Mario can be conscious of Naples: Mario is the subject of the thought, and Naples its object. Mario, however, can also be conscious of himself*: in this case, Mario is both subject and object of the thought. Even though there is one single entity, and the subject and object of the thought are the same thing, it is possible to draw a conceptual distinction between Mario's ability as object of thought and Mario's ability as subject of thought: namely, the concepts of self-as-subject and self-as-object. James (1890) distinguishes between *I* and *me,* so that we can tell the difference between "I am self-conscious that I think that *p*" and "I am self-conscious that me thinks that *p*". While the former refers to the self-as-subject, "me" refers to the self-as-object:

> Corresponding to these two concepts, or conceptions, of self, there would presumably be two distinct modes of presentation under which a person may be conscious of herself. She may be conscious of herself under the "I" description or under the "me" description. Thus, my state of self-consciousness may employ either the "I" mode of presentation or the "me" mode of presentation. […] In the latter case, there is a sort of "conceptual distance" between the thing that does the thinking and the thing being thought about. Although I am thinking of myself, I am not thinking of myself as the thing that does the thinking. By contrast, in the former case, I am thinking of myself precisely as the thing that is therewith doing the thinking. (Kriegel [2007])

The self as subject may be thus interpreted as the thing that does the thinking, whereas the consciousness of oneself as subject is the consciousness of oneself as the thing doing the thinking.

It seems clear that the first-person perspective and the consciousness of self as subject are two interdependent features, the one being the condition of the other and *vice versa.* The subject's manifestation of strong first-

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

person phenomena – one's ability to attribute first-person reference to herself – is based on the subject's possibility of being the consciousness of herself as the thing doing the thinking, and the subject's consciousness of self-as-subject cannot be gained unless the subject exhibits a manifestation of strong first-person phenomena.

As has just been said, such two features defining the notion of basic self-consciousness are grounded on a few epistemic and semantic peculiarities in the ability to use the term or concept *I* in *de se* or *I-thoughts*: the essential indexicality and the immunity to error through misidentification. These will be discussed in parts 2 and 3.

2. Indexicality

In the following passage Castañeda (1990, 736) points out the importance of indexicality not only for a philosophical-linguistic reflection:

> Indexicality is certainly a linguistic phenomenon, but one which reveals something of great importance about our intentional representations of the components of the world. Indexicality is […] the network, and the structure, of thought contents we express when we use indicators (demonstratives, the main simple tenses, personal pronouns) to refer to items of experience as experienced: perceived *this*'s and *that*'s; experienced *now*'s; perceived *here*'s and *there*'s; *you*'s being addressed; experienced experiencing *I*'s. The tokens of these expressions so used are singular terms, almost like logicians' individual constants, except that they have a content, an irreducible thought content. Of course, experiences can go on without being expressed. Hence, the primary indexical mechanisms are internal mechanisms of representation of the appropriate type of experienced item. Indexicality is the backbone of experience. Indexical reference is the procedure through which a thinker picks out pieces or aspects of reality for experiential confrontation.

As has already been said, the I*-thought allows an individual to refer to herself as herself* without a need for third-person referential devices, such as names, descriptions, or demonstratives to identify herself: the *I* employed in a self-conscious or I*-thought is essentiality indexical[4]; as

[4] Following Kaplan (1989) and Perry's (1997, 594) classic approaches "a defining feature of indexicals is that the meanings of these words fix the designation of specific utterances of them in terms of facts about these specific utterances. The facts that the meaning of a particular indexical deems relevant are the contextual facts for particular uses of it". Kaplan and Perry have detected several difficulties in the Fregean approach to indexical terms, in particular with the notion of *sense*.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

such, it necessarily involves information indexed to the context and, more specifically, to the thinker who has produced the thought. More specifically, (a) "I" is a singular term/concept, i.e., a term with a single individual as its reference; (b) this term is governed by the token-reflexive rule, whereby every token of "I" refers to the subject who has produced or used it, either mentally or linguistically; (c) with the information available in context, and once established the circumstances of evaluation, this rule is *prima facie* sufficient to determine its reference[5]. More importantly, the indexical information about oneself based on the use of the term/concept "I" cannot be reduced to non-indexical information; for this reason, indexicality is *essential*. Two well-known examples by Perry (1977) describe the matter at issue.

The first example is about a fictional character named Rudolf Lingens: "An amnesiac, Rudolf Lingens, is lost in the Stanford library. He reads a number of things in the library, including a biography of himself, and a detailed account of the library in which he is lost. […] He still won't know who he is, and where he is, no matter how much knowledge he piles up, until that moment when he is ready to say, *This* place is aisle five, floor six, of Main Library, Stanford. *I* am Rudolf Lingens". Amnesiac Rudolf Lingens can gather all sorts of information about himself by reading the books in the Library, and yet no such information can provide him with the missing conceptual tool he needs to link the information with himself. In other words, there is no logical connection between third-person descriptive information, no matter how detailed, and a first-person grasp of oneself through the use of *I*.

The second example regards indexical judgements (beliefs and desires), which are crucial to explain and predict the motivating action: «I

Kaplan (1989) distinguishes between *character* and *content*, the former being the linguistic meaning fixed by linguistic convention, and the latter determined according to the context associated with at least one agent, time, location, and possible world. On the other hand, Kaplan distinguishes between *pure indexicals* (for example, "I", "now", "here") and *impure indexicals* (also called *true demonstratives* or *deictic terms*: e.g., "this", "that"). In the latter case, the determination of reference is not only based on the rule associated but also on the mediation of the subject's intention.

[5] This approach to indexicality has been acknowledged by major scholars, cf. Shoemaker (1968, 91), Peacocke (1983, 133–9), Rovane (1987, 147), Campbell (1994, 73), Kaplan (1989, 493). Nonetheless, their positions are not entirely uniform – cf. Castañeda (1983; 1989), Kapitan (2001; 2008), and de Gaynesford (2006) for an analysis of the debate.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch» (Perry, 1979).

In the example, Perry's thoughts (a) "the shopper with the torn sack is making a mess" and (b) "I am making a mess" refer to the same individual; the point is that two intrinsically different kinds of self-reference are at play here. In the former, self-reference is external, and only available in the third person: Mario can refer to an object by using a name, a definite description or demonstrative, and the object he is referring to might be himself; there is no difference between this kind of self-reference and the reference made to an object that is different from oneself. An external self-reference can occur without the subject's realizing he is referring to himself, as in the first thought formulated by Perry, where he does not realize he is the very shopper with the torn sack who is making a mess. This recalls the well-known particular episode told by Mach (1914): "Not long ago after a trying railway journey by night, when I was very tired, I got into an omnibus, just as another man appeared at the other end. 'What a shabby pedagogue that is, that has just entered', thought I. It was myself: opposite me hung a large mirror. The physiognomy of my class, accordingly, was better known to me than my own".

On the other hand, the internal self-reference expressed by the second thought (b) "I am making a mess" produces an authentic I-thought that is only accessible from a first-person perspective because it is based on the use of *I*. After realizing that he is the shopper at issue, Perry produces a new thought, which he terms *locating belief*, based on the use of the essential mental indexical *I*. This entails that the ascription of an authentic I-thought to oneself cannot be achieved without the concept *I*, as there is no way to think an I-thought other than through indexical reference; hence, I-thoughts are irreducible to the other kind of non-indexical thoughts.

Accordingly, by articulating an I-thought in a propositional way, an I-thought will contain a content whose subjective essential indexical reference is expressible in the natural language by the personal pronoun *I*. This thought can be reported in a direct or indirect form: while the former is in *oratio recta* and reports the above-mentioned example as in (1) "I am making a mess" (thought by Perry), in the indirect form the thought can be expressed in *oratio obliqua,* and the report will be in the third person: (2) "Perry thinks that he himself is making a mess". In turn, this sentence can

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

be interpreted as the report of yet another thought still: (3) "Perry thinks that Perry is making a mess". Obviously, it is possible to employ a definite description, "the φ" that picks out Perry uniquely, as in this example: (4) "Perry thinks that the author of *The Essential Indexical* is making a mess".

The thought expressed in (1) is neither equivalent to (3) nor to (4). With cases (3) and (4), Perry might be amnesiac and remember neither his name nor his being the author of *The Essential Indexical*. It is only in (1) that an authentic *I-thought* is present: the subject who thinks the thought "I am making a mess", provided that she knows the rule associated with *I*, cannot use it without realizing that she is referring to nobody but herself. Although (1) and (3) or (4) are not equivalent, *prima facie* (2) seems to be a report of both. To capture the difference, Castañeda employs two different uses of the third person pronouns in sentences in *oratio obliqua*. In the first case, to make (2) equivalent to (1), the pronoun is to be used in an *indirect reflexive* manner (Anscombe, 1975) or as a *quasi-indicator*, the above-mentioned artificial pronoun ("she*, he*, it*") introduced by Castañeda (1966; 1967; 1968) to attribute a first-person essential indexical use from a third-person angle: (2.1) "Perry thinks that he* himself is making a mess". The *quasi-indicator* "he*" in the example is used as an anaphora, and its reference is not determined directly but only through the propositional attitude subject. In the other case, assuming that Perry is amnesiac, to make (2) equivalent to (3) "he" will not be employed as a quasi-indicator but as a simple indexical; thus, (2) is the report of Perry's belief that someone else in the context (named Perry too) is making a mess: Perry has not realized that it is he* himself who is doing that.

This is the difference introduced by Castañeda (1999, 256) to distinguish between what he labels external and internal reflexivity of self-consciousness: only the former can produce an authentic I-thought describable in the third person through the quasi-indicator:

> There is the external reflexivity of ONE referring to ONEself, as when shaving ONE accidentally cuts ONEself, rather than another. […] Like the reflexivity involved in cutting oneself, the external reflexivity of reference to oneself can be unintentionally and unwittingly executed. Thus a forgetful painter may think that the painter of a certain picture is a very good painter without realizing that he himself painted the picture. There is, on the other hand, the internal reflexivity of ONE referring to something, whatever it may be, as oneself. The internal reflexivity is the peculiar core of self-consciousness.

In this manner, token-reflexive expressions such as first-person pronouns and quasi-indicators are essential indexicals: they can be neither eliminated nor replaced by a name, description, or demonstrative without

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

losing the content expressed by the sentences/thoughts that contain them: to refer to (to think of) oneself *qua* oneself, the subject has to use the essential indexical *I*. As Shoemaker (1968, 560) condenses, «there is no description at all which is free of token-reflexive expressions and which can be substituted for 'I'; no matter how detailed a token-reflexive-free description of a person is, [...] it cannot possibly entail that I am that person».

In other words, essentiality it lies in the fact that the use of *I* is based on no cognitive mediation, but is precisely indispensable or essential to form I-thoughts. This entails that an identifying description is neither necessary nor sufficient condition to the self-reference of the self-conscious subject. In one case, an amnesiac subject locked in a completely dark room is able to use *I* in order to refer to herself even though she cannot apply third-person descriptions to herself. Castañeda clarifies (1999, 268-9):

> There is no third-person special characteristic that one has to think that one possesses in order to think of oneself as *I*. Certainly, one *qua* I does not classify oneself as a self, a person, or a thinker – let alone as a human being, female, or whatever is true of all entities capable of *self* consciousness. To illustrate, a small child at about the age of two can make perfect first-person references fully lacking knowledge involving those categories. […] There is just no criterion one can apply to determine whether one is an *I* or not. One simply is an I. This primitive fact is primitively and immediately apprehended by a thinker who is an *I*.

In the other case, as has already been said with Shoemaker and suggested by Perry's example of amnesiac Rudolf Lingens, no matter how detailed a third-person description about a certain subject, the latter can grasp it without realizing she is the subject the description refers to.

The self-reference expressed in linguistic or mental self-ascriptions such as (2.1) is referred to as *de se* (i.e., of oneself), an expression used in Lewis (1979) and Castañeda's works, and is semantically unique: the linguistic or mental self-reference is irreducible to either *de dicto* or *de re* linguistic or mental reference (for the above considerations, the truth conditions of the *de se* "Perry thinks that he* himself is making a mess" are different from both *de dicto* "Perry thinks that Perry is making a mess" and *de re* "Perry thinks, of Perry, that he is making a mess"). If *de se thoughts* are mental states reported in *de se* reports, then they are irreducible to mental states reported in *de dicto* and *de re* reports. In this way, *I-thoughts* make up an irreducible class of mental phenomena.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Immunity to Error through Misidentification

In a well-known passage, Wittgenstein (1958, 66-7) introduces his philosophico-linguistic analysis of the grammatical rule of the term I, where he identifies two types of uses, i.e., "I" used as object ("I have grown six inches") and "I" used as subject ("I have toothache"): "One can point to the difference between these two categories by saying: The cases of the first category involve the recognition of a particular person, and there is in these cases the possibility of an error... On the other hand there is no question of recognizing a person when I say I have toothache. To ask 'are you sure it's you who have pains?' would be nonsensical".

This passage should be taken as part of the philosophical framework articulated by Wittgenstein since the 1930s according to some theses to be regarded as the background for the analyses of the two uses of *I*. While the *I* used as object performs a referential function relative to one's body and physical features in general, the *I* used as subject apparently regards mental states and processes involving no subject identification[6].

[6] From a Wittgensteinian angle, the *I used as subject* performs no referential function: according to this thesis – supported by Geach (1957), Hacker (1972), Anscombe (1975) – it is only our inclination to assume that a linguistic term has a meaning only if it stands for an object that makes us believe that the *I used as subject* denotes the thinking subject, mind, soul, etc. (cf. Sluga 1996, Wright 1998). In this way Wittgenstein (1958, 43) starts from the analysis of language and the use of the *I as subject* to dissolve any question on the nature of the ego in an anti-metaphysical key. Philosophical inquiry must investigate only the *grammars* of the mentalistic terms used and no metaphysical distinction between the mental and the physical should follow from the distinction between propositions describing facts of the world and propositions describing psychological experiences. It is necessary to analyse the uses and related grammars of terms such as *thinking*, *meaning*, *wishing* because the investigation "rids us of the temptation to look for a peculiar act of thinking, independent of the act of expressing our thoughts, and stowed away in some peculiar medium". Thinking is using signs according to rules, and philosophical difficulties may arise only from the misleading use of language which leads us to look for something that might correspond to a noun. This may be the case in the use of the *I as subject*: the referential thesis according to which the use of a sign is based on its relation with the object – strongly criticized when taken as the sole basis to explain the semantics of the language, along with the proper consideration that some uses of the *I* do not denote physical properties – lead to false Cartesian metaphysical conclusions: "We feel then that in the cases in which "I" is used as subject, we don't use it because we recognize a particular person by his bodily characteristics; and this creates the illusion that we use this word to refer to something bodiless,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Similarly, Strawson (1966, 165) argues that in the self-ascription of a mental state (e.g., "I'm hungry") a subject of experiences uses the term "I" without any identification criteria: "It would make no sense to think or say: This inner experience is occurring, but is it occurring to me? (This feeling is anger; but is it I who am feeling it?)". More precisely, Strawson refers to *criterionless self-ascription*: "When a man (a subject of experience) ascribes a current or directly remembered state of consciousness to himself, no use whatever of any criteria of personal identity is required to justify his use of the pronoun 'I' to refer to the subject of that experience". On the other hand – and in contrast to Wittgenstein – the absence of an identification device does not entail that the use of *I* performs no referential function whatsoever.

Particular judgments with a first-person reference (e.g., "I have pain") display what Shoemaker (1968, 565) defines *self-reference without identification*: "My use of the word 'I' as the subject of my statement is not due to my having identified as myself something of which I know, or believe, or wish to say, that the predicate of my statement applies to it". The self-ascription of the thoughts on which the self-consciousness is based regards the consciousness of oneself *qua* subject – i.e., as the subject of every thought or mental state – rather than as the object based on the previous identification component.

In other words, in the self-ascriptions of mental properties the self-reference underlying particular self-conscious forms occur without any inference from conceptual properties ascribable to the subject: there is no previous identification of something as its own self owing to properties that can be ascribed to that same something. Due to the absence of any identification component, particular singular judgments involving the self-ascription of mental (and physical, as will be seen) properties are *immune to error through misidentification* relative to the first-person pronoun (IEM). The subject formulating such judgments in given epistemic contexts cannot be mistaken as to whether it is he who is attributing a particular mental property to his own self:

> (…) to say that a statement 'a is Φ' is subject to error through misidentification relative to the term 'a' means that the following is possible: the speaker knows some particular thing to be Φ, but makes the mistake of asserting 'a is Φ' because, and only because, he mistakenly thinks that the thing he knows to be Φ is what 'a' refers to. The statement

---

which however, has its seat in our body. In fact this seem to be the real ego, the one of which it was said, 'Cogito ergo sum'" (Wittgenstein 1958, 69).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> 'I feel pain' is not subject to error through misidentification relative to 'I': it cannot happen that I am mistaken in saying 'I feel pain' because, although I do know of someone that feels pain, I am mistaken in thinking that person to be myself. (Shoemaker [1968, 557])

To make this passage clear, we might consider an example of identification-dependent thought, e.g., Mario's thought that his neighbour is a nice person. Following Evans (1982) and Kriegel (2007), the structure of this thought consists of an identification component and a predication component, which can be explicated by Mario's first-person perspective as follows: "my neighbour [identification component] is a person who smiles at me every day and the person who smiles at me every day is a nice person [predication component]". Here, two types of errors are possible. Mario can be mistaken as to the predicational component, i.e., that his neighbour is a nice person: for example, later on Mario finds out that his neighbour's tendency to smile is nothing but a cynical strategy to have him consent to cut the trees in the garden. Mario can also be mistaken as to the identificational component – i.e., as to the person who is his neighbour – and, for example, get confused and mistake the mailman for his neighbour.

On the other hand, there is a class of self-ascriptions involving no identification-dependent thoughts; as such, it is not subject to error through misidentification. Shoemaker actually introduces finer distinctions – i.e., between *de facto* and *logical IEM* on the one hand, and between *circumstantial* and *absolute IEM* on the other – pondering a class of psychological predicates involved in those self-ascriptions which are absolute *IEM* as opposed to physical predicates' self-ascriptions. For example, when Mario sees a large number of hands in the mirror, his judgment "I have a dirty hand" can be subject to two types of errors: Mario can be mistaken as to the fact that his hand is dirty, and about the owner of the hand. Instead, if Mario thinks that he is happier than last year, he cannot be mistaken as to whether it is he who is attributing a particular mental property to his own self. Therefore, for Shoemaker as well as Wittgenstein, there is a class of self-ascriptions that is immune to error based on the misidentification of the subjective reference, no identification component being involved.

Shoemaker (1968, 565) examines the kinds of psychological predicates involved in such self-ascriptions: "There is an important and central class of psychological predicates, let us call them *P** predicates, each of which can be known to be instantiated in such a way that knowing it to be instantiated in that way is equivalent to knowing it to be instantiated in oneself". For instance, the judgment "I have pain" is IEM because the way in which the predicate is expressed ("there is pain"), that is, based on our

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

own subjective experience, will suffice to realize that it is ascribed to ourselves ("I have pain"). It is in this particular sense that "there is pain" is tantamount to "I have pain".

In turn, Evans goes beyond the terms of the matter as suggested by Wittgenstein and, to some extent, by Shoemaker. In particular self-ascriptions, the self-reference is direct and unmediated: this, as Evans notes, is *identification-free self-reference*. More to the point, moving from the self-ascription of properties that are not only mental but also physical, the author discloses his approach: judgments are IEM when they result from the connection between the information acquired in the first person and the information justification, as opposed to *identification-dependent* judgments involved in the ordinary perception of external objects. The IEM feature does not depend on the kind of predicate involved in the self-ascription but on the epistemic and justification ground on which the subject produces such judgments (cf. Wright 1998, 19) in a context where – from Strawson's lesson onward – the subject is conceived as a spatio-temporally located object.

As Wright (1998, 19) points out: "the ground has to be such that in the event that the statement in question is somehow defeated, it cannot survive as a ground for the corresponding existential generalization". Going back to the example examined before, "there is pain" is tantamount to "I have pain"; if the second judgment is defeated – for instance, if it is not true that "I have pain" – the kind of access to the information through the subject's first-person mental dimension on which the judgment has been made does not allow us to maintain the first judgment, "there is pain", i.e., the corresponding existential generalization: if I am not the one who has the pain, then there cannot be anybody else with the pain. Instead, in the judgments displaying uses of *I* as object, the error through misidentification is always possible. In such cases, the corresponding existential generalization survives: if Mario is mistaken as to who is the owner of the hand he is seeing in the mirror, and he realizes he is not the owner of the hand, the corresponding existential generalization "someone has the dirty hand" survives.

In this way, Evans (1982, 220) contends that a judgment such as 'I am F' is identification-free unless it corresponds to the inferential conclusion drawn from the two premises, i.e., 'a is F' (predication component) and 'I am a' (identification component). Such a judgment is based on the unmediated self-ascription of properties through the introspective consciousness (as is the case with mental properties) or proprioception (as with physical properties). For example, according to our general capacity to perceive bodies and to our sense of proprioception, of balance, of heat

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and cold, and of pressure, the kind of information generated by each of these modes of perception seems to give rise to immune to error through misidentification judgments: "None of the following utterances appears to make sense when the first component expresses knowledge gained in the appropriate way: 'Someone's legs are crossed, but is it my legs that are crossed?' ".

Peacocke's (2008) strategy, in turn, consists in associating IEM properties with more fundamental characterizations. By specifying the manner and circumstances in and under which an IEM judgment relative to the occurrence of a particular concept is formed, Peacocke pushes into the background the characterizations of the functional rule of first-person concepts highlighted by Evans. In addition, he contends that the best explanation for the phenomenon of judgments formed according to the self-attributions of mental and physical properties relies on the basic reference rule for which "any use of I refers to the thinker of the thought in which it occurs".

In actual fact, the author has already focused on the first-person concept's possession conditions, which are grounded on its capacity to be employed provided that one is conscious of a particular mental state. More precisely, Peacocke (1999) distinguishes between representationally dependent and independent uses of the first-person concept to define what he terms *delta account*. The point at issue here is primarily epistemological: it concerns the philosophical branch of self-knowledge as well as the possibility of forming beliefs relative to the self-ascription of mental and physical properties. While the representationally dependent use of the first-person concept is based on the fact that the subject is represented in the content of the judgment, in the representationally independent use of the first-person concept the very occurrence of a particular experience (namely visual, its content in Peacocke's example being "I see the phone is on the table") determines the reason why the subject is justified in making a judgment about herself, without the thinking subject being represented in the judgment: "the explanation is just the occurrence of the experience itself to its subject. Nor does any thought or representation of herself as the subject of the experience enter her reasons for her judgment". *Mutatis mutandis* Recanati (2007; 2009) enacts a similar strategy through a philosophical analysis of the distinction between *de re* and *de se thoughts*, as maintained by Chisholm (1976; 1981), Lewis, and Perry.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Conclusion

To sum up, the basic capacity for self-consciousness discussed in these pages depends 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.

The analysis of the form of the concept *I* is intertwined with several epistemic and metaphysical questions. In general, it should be highlighted that the absence of an identification component does not imply that the *I* doesn't perform a referential function, nor that it necessarily involves a specific metaphysical thesis on the nature of the self-conscious subject. As a matter of fact, the *I-thoughts* self-reference features have been supported by both a materialist conception regarding the self-conscious subject as a bodily object – for example, by Strawson and Evans – and a different metaphysical framework, as in Wittgenstein's eliminativist thesis or in Kant's exclusion thesis (cf. Forgione [2013; 2015]).

Before closing the chapter, it is worth touching upon, at least, another very important kind of self-consciousness first analysed by the phenomenological tradition: even though phenomenologists disagree on methods and issues, the major figures in phenomenology (above all, Husserl and Sartre) support the argument that a minimal form of self-consciousness is a constant intrinsic feature to whichever conscious experience. This form of self-consciousness is called "pre-reflective" and, along with the basic self-consciousness, is a very important point at issue in the philosophical debate on self-consciousness. In Gallagher and Zahavi's words (2008, 46), "[Pre-reflective self-consciousness] is not thematic or attentive or voluntarily brought about; rather it is tacit, and very importantly, thoroughly nonobservational (that is, it is not a kind of introspective observation of myself) and non-objectifying (that is, it does not turn my experience into a perceived or observed object). I can, of course, reflect on and attend to my experience, I can make it the theme or object of my attention, but prior to reflecting on it, I wasn't 'mind- or self-blind'. The experience was already present to me, it was already something *for me*, and in that sense it counts as being pre-reflectively conscious". Similar views are present in analytic philosophy as well[7]. It

[7] Two examples are provided. The first is Flanagan (1992, 194): "all subjective experience is self-conscious in the weak sense that there is something it is like for the subject to have that experience. This involves a sense that the experience is the subject's experience, that it happens to her, occurs in her stream". The other example is Frankfurt (1988, 162): "What would it be like to be conscious of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

follows that self-consciousness presents two modes of existence, a pre-reflective and a reflective one: needless to say, and in referring to Zahavi (2005) in order to examine the issue in depth, one of the most interesting philosophical questions concerns the very relation between these two utterly different forms of self-consciousness.

References

Almog, J., Perry, J., Wettstein, H. (eds.) (1989), *Themes from Kaplan*. Oxford: Oxford U.P.

Anscombe, G. E. M. (1975), The First Person, in *Cassam* (1994) (pp. 140-159).

Baker, L. R. (2000), *Persons and Bodies: A Constitution View*. Cambridge MA: Cambridge U.P.

Baker (2013), *Naturalism and the First-person Perspective*. New York: Oxford U.P.

Campbell, J. (1994), *Past, Space and Self*. Cambridge MA: MIT Press.

Cassam, Q. (ed.) (1994), *Self-Knowledge*. Oxford: Oxford U.P.

Castañeda, H.-N. (1966), 'He': A Study in the Logic of Self-Consciousness, in Id. (1999) (pp. 35-60).

—. (1967), Indicators and Quasi-Indicators. In *American Philosophical Quarterly* 4 (pp. 85–100).

—. (1968), On the Phenomeno-Logic of the I, in Id. (1999) (pp. 89-95).

—. (1983), Reply to John Perry: Meaning, Belief, and Reference. In Tomberlin J. (ed.), (pp. 313-328).

---

something without being aware of this consciousness? It would mean having an experience with no awareness whatever of its occurrence. This would be, precisely, a case of unconscious experience. It appears, then, that being conscious is identical with being self-conscious. Consciousness is self-consciousness. The claim that waking consciousness is self-consciousness does not mean that consciousness is invariably dual in the sense that every instance of it involves both a primary awareness and another instance of consciousness which is somehow distinct and separable from the first and which has the first as its object. That would threaten an intolerably infinite proliferation of instances of consciousness. Rather, the self-consciousness in question is a sort of immanent reflexivity by virtue of which every instance of being conscious grasps not only that of which it is an awareness but also the awareness of it. It is like a source of light which, in addition to illuminating whatever other things fall within its scope, renders itself visible as well".

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. (1989), *Thinking, Language, and Experience*. Minneapolis: University of Minnesota Press.
—. (1990), Indexicality: The Transparent Subjective Mechanism for Encountering A World, "Noûs" 5(24), pp. 735-749.
—. (1999), The Phenomeno-Logic of the I. Essay on Self-consciousness. Bloomington: Indiana U.P..
Chisholm, R. (1976), *Person and Object: A Metaphysical Study*. Chicago and La Salle: Open Court.
—. (1981), *The First Person*. Minneapolis: University of Minnesota Press.
de Gaynesford, M. (2006), *I: the Meaning of the First Person Term*. Oxford: Oxford U.P.
Evans, G. (1982), *The Varieties of Reference*. Oxford: Oxford U.P.
Flanagan, O. (1992), *Consciousness Reconsidered*. Cambridge MA: Cambridge U.P.
Forgione, L. (2013), Kant and I as subject, in S.Bacin, A. Ferrarin, C. La Rocca, M. Ruffing (eds.), Kant und die Philosophie in Weltbürgerlicher Absicht. Akten des XI. Kant-Kongresses 2010, de Gruyter, Berlin.
—. (2015), Kant and the Problem of Self-Identification, "Organon F" 22 (2), pp. 178-197.
Frankfurt, H. (1988), The importance of what we care about: philosophical essays. Cambridge MA: Cambridge U.P.
Gallagher, S., Zahavi, D. (2007), The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science. London: Routledge.
Geach, P. T. (1957), *Mental Acts: Their Content and Their Objects*. London: Routledge.
Gertler, B. (2010), *Self-knowledge*. London: Routledge.
Hacker, P. M. S. (1972), *Insight and Illusion*. Oxford: Clarendon Press.
James, W. (1890), *The Principles of Psychology*. New York: Cosimo, 2007.
Kapitan, T. (2001), Indexical Identification: A Perspectival Account, "Philosophical Psychology" 14 (3), pp. 293-312.
—. (2008), Perry, Castañeda, and 'I', The XVIII edition of the Inter-University Workshop on Philosophy and Cognitive Science, http://fs-morente.filos.ucm.es/actividades/2008/conference/papers/Kapitan.pdf
Kaplan, D. (1989), Demonstratives and Afterthoughts, in Almog J., Perry J., Wettstein H. (eds) (1989), pp. 481–614.
Kriegel, U. (2003), Consciousness as Intransitive Self-Consciousness: Two Views and an Argument, "Canadian Journal of Philosophy", 33, pp. 103-132.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. (2007), Self-Consciousness, "Internet Encyclopedia of Philosophy", http://www.iep.utm.edu/self-con/

Lewis, D. (1979), Attitudes De Dicto and De Se, in Id., Philosophical Papers, Oxford U.P., Oxford, 1983, pp. 133-159.

Mach, E. (1914), *The Analysis of Sensations*. Chicago and London: Open Court.

Matthews, G.B. (1992), *Thought's Ego in Augustine and Descartes*. Ithaca, NY: Cornell U.P.

Peacocke, C. (1983), *Sense and Content*. Oxford: Oxford U.P.

—. (1999), *Being Known*. New York: Oxford U.P.

—. (2008), *Truly Understood*. Oxford: Oxford U.P.

Perry, J. (1977), Frege on Demonstratives, "Philosophical Review" 86 (4), pp. 474-497.

—. (1979), The Problem of the Essential Indexical, in Cassam Q. (ed) (1994), pp. 167–83.

—. (1997), Indexicals and Demonstratives, in Hale B., Wright C. (eds), A Companion to the Philosophy of Language, Blackwell, Oxford, pp. 586-612.

Recanati, F. (2007), *Perspectival Thought. A Plea for (Moderate) Relativism*. New York: Oxford U.P.

—. (2009), De re and De se, "Dialectica", 63, pp. 249-269.

Rovane, C. (1987), The Epistemology of First-Person Reference, "Journal of Philosophy", 84, pp. 147–167.

Sluga, H. (1996), Whose house is that? Wittgenstein on the self, in H. Sluga, D. G. Stern (eds.), The Cambridge Companion to Wittgenstein, Cambridge U.P. Cambridge MA.

Shoemaker, S. S. (1968), Self-reference and self-awareness, "The journal of philosophy", 65(19), pp. 555-567.

Strawson, P.F. (1966), *The Bounds of Sense. An Essay on Kant's Critique of Pure Reason*. London: Methuen.

Wittgenstein, L. (1958), *The Blue and the Brown Books*. Oxford: Blackwell.

Wright, C. (1998), Self-knowledge. The Wittgensteinian Legacy, in Wright C., Smith B.C., Macdonald C. (eds), Knowing Our Own Minds, Oxford U.P., Oxford, pp.12-45.

Zahavi, D. (2005), *Subjectivity and Selfhood: Investigating the First-person Perspective*. Cambridge MA: MIT Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART II:

ODORS, COLORS AND VISION

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER FOUR

ENACTIVISM'S LAST BREATHS

BENJAMIN D. YOUNG

Olfactory perception provides a promising test case for enactivism, since smelling involves actively sampling our surrounding environment by sniffing. Smelling deploys implicit skillful knowledge of how our movement and the airflow around us yield olfactory experiences. The hybrid nature of olfactory experience makes it an ideal test case for enactivism with its esteem for touch and theoretical roots in vision. Olfaction is like vision in facilitating the perception of distal objects, yet it requires us to breath in and physically contact the sensory object in a manner similar to touch. The paper offers an analysis of the central theoretical components of enactivism, whose soundness and empirical viability are tested using the empirical literature on sniffing. It will be shown that even if sniffing is an essential component of olfaction, the motoric component is not necessary for perceiving smells, which is contrary to the most crucial tenet of enactivism. Thus, the paper concludes that the theory cannot account for olfactory perception.

Introduction

The behavioral rhythm of breathing, which is under volitional control, modulates our inhalation and exhalation patterns in accordance with our internal states and surrounding environment. Even our breathing patterns are a form of interaction with our environment that facilitate olfactory perception and that can be utilized in testing enactivism. Enactivism argues that perception is constituted by the implicit deployment of skill based sensorimotor knowledge, which structures the nature of our perception of external objects in the environment (Noë [2001, 2002, 2005], O'Regan [1992]; O'Regan and Noë [2001, 2002]. Consequently, Alva Noë (2004, 2007, 2008, 2009), the most vocal proponent of enactivism, suggests that touch would serve as a better theoretical

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

launchpad for theorizing about the nature of perception than vision. Yet, he, like many others, concentrates upon vision with the promise that the theory should extrapolate to the other senses.

However, there are good reasons to doubt that the theory can accommodate the nature of taste (Gray and Tanesini [2010]), making its extrapolation to other modalities dubious, especially when compounded with the evidence that it is an inadequate account even for its designed domain of visual perception (Lycan [2006]; Prinz [2006]). Enactivism is unlikely to serve as a comprehensive theory of all forms of perception, yet our perception of smells might offer it a last bastion of hope. Rather than focus on all aspects of breathing the paper concentrates on the role of sniffing in olfactory perception. *Prima facie* olfactory perception provides a promising test case for enactivism, since smelling requires actively sampling our surroundings by sniffing. Olfactory perception depends upon the movement of chemical compounds through the nose.

The hybrid nature of the olfactory experience makes it an ideal test case for enactivism with its esteem for touch and theoretical roots in vision. Olfaction is similar to vision in enabling the distal perception of object. We perceive distal smells within the environment, yet physically contacting the sensory object in a manner similar to touch is required to transduce the smell stimuli. A survey of the empirical research on sniffing will be employed to both generate an analysis of the theoretical components of enactivism and test the soundness of the theory.

Airflow through the nose is required for normal cases of olfactory perception whether it be orthonasal, originating from the front of the nostrils, or retronasal, passing up from the throat through the back of the nose. However, for the purposes of this paper, only orthonasal perception will be considered, as the inhalation of odorant laced air as controlled by sniffing is the focus of this paper. Moreover, rodents, who serve as the primary animal models of olfactory perception related to sniffing, only have orthonasal olfactory perception. Thus, olfactory perception as initialized at the nostrils and flowing back to the olfactory epithelium will be the bases for sniff testing the enactive approach. It will be shown that though sniffing is at times an essential component of the olfactory percept, the motoric component is not necessary for perceiving smells. Since this is contrary to the crucial tenet of enactivism, the conclusion argued for is that the theory cannot account for olfactory perception.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

1. Sniffing

The average sniff lasts 1.6 seconds. During the initial phase of sniffing we modulate the volume of airflow, pressure of airflow, and sampling rates. Additionally, towards the middle to end of a sniff we can detect the presence of an odor, as well as identify its olfactory quality (what it smells like) and valence (reviewed in Olofsson [2014]). The sniff sequence can be segmented into multiple stages. The initial sniff onset brings the stimulus into the nasal cavity and lasts 200 ms. Within 150-300 ms of stimulus presentation sniffing is modulated in accordance with the concentration, intensity, and valence of the odorant. Additionally, within 150 ms of sniff onset we modify our sniff response in accordance with the olfactory valence of the stimulus. Furthermore, encoding the olfactory properties of the odor occurs during a 500 ms period following the initial 200 ms of sniff onset. Only after 800 ms of sniff onset do we consciously detect the odorant. Identification of olfactory quality and odor valence follows at intervals of approximately 1000 ms and 1100-1200 respectively (reviewed in Olofsson [2014]).

What will be of interest in testing enactivism is that the behavioral modulation of our nostrils and breathing patterns, a paradigm of motoric action, occurs before we consciously report experiencing a smell. The behavioral modulation of each sniff serves not only to deliver the odorant to the olfactory epithelium, but it also plays a role in determining our perception of odor intensity and concentration. Our experience of smell is modulated by our sniffing behavior, which tracks the olfactory properties of the odorant below the level of conscious awareness.

Not only does sniffing facilitate our perception of a smell's valence, intensity, and concentration, it enables the localization of olfactory objects by actively exploring the external smellscape (Porter et al. [2007]). Having two nostrils serves a greater function than mere aesthetic symmetry, the two nostrils form distinct percepts of the olfactory environment and based on their differences in anatomical size and volume of airflow (Sobel et al. [2000]; Zhou and Chen [2009]) we are able to track olfactory objects across time and throughout an environment (Porter et al. [2007]).

The role of sniffing in generating the olfactory percept provides a promising line of evidence in favor of testing enactivism. But sniffing serves non-perceptual functions as well. Amongst rats, sniff rates are modulated based on social hierarchy. Males of lesser ranks will decrease their sniffing rates around those of a greater stature, because vigorous sniffing can be interpreted as a sign of aggression (Wesson [2013]). There is no evidence for the same social analogue in humans, but experimental

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

research demonstrates that we mimic the sniffing behavior of those around us. In these instances, sniffing might be a partial mechanism for directing shared attention towards olfactory objects (Arzi et al. [2014]). Sniffing plays a role in olfactory perception, but it might serve other purposes than just facilitating smelling (Galef [2013]). In what follows, the key target of this paper, actionism will be introduced, and its key tenets will be identified, clarified and tested, using what we know about the role of sniffing in olfactory perception.

2. Actionism

There is little doubt that sensorimotor contingencies play a role in our perception of the world, but enactivism endorses a non-trivial and strong constitutive relationship between them. Central to the theory is the claim that perceiving objects in the environment is only possible through the existence, knowledge and implicit deployment of sensorimotor contingencies. To perceive objects in the environment one must tacitly understand how one's movement would affect the sensory properties of external perceptible entities. The target of the paper is Noë's most recent incarnation of enactivism, actionism. His current theory is in keeping with the past versions, but with a more explicit formulation of the central claim that to perceive is a skillful act of knowing how movement affects the perception of an object's sensory properties. Actionism expands this idea into two separate conditions:

> (i) Movement-dependence: movements of the body manifestly control the character of the relation to the object or quality (ii) Object-dependence: movements or other changes in the object manifestly control the character of the relation to the object or quality. (Noë [2009], p. 476)

The first claim is consistent with previous theories (Noë [2001, 2002, 2005, 2004, 2007]; O'Regan and Noë [2001, 2002]), while the second is a fresh addition. The further addendum of object dependence seems like a banal augmentation that most perceptual scientists would agree on without much fuss. Surely, our ability to track perceptual object throughout their shifting mereology and changes in spatiotemporal properties depends on our knowledge of how objects move throughout an environment in a stable manner based on perceptual constancies. What remains the non-trivial and contentious claim is that of movement-dependence to which we now turn our attention.

Actionism has four foundational theoretical tenets. First, the theory tacitly endorses a model of direct perception. Second, actionism argues

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

that the transition from sensation to perception requires sensorimotor knowledge. Third, in keeping with its enactivist roots actionism maintains the constitutive claim that perception requires the possession of implicit knowledge and the skillful deployment of the aforementioned sensorimotor contingencies. Lastly, actionism attempts to explain the perceptual presence of three-dimensional objects using our knowledge of sensorimotor contingencies that determine perceptual constancies. To test the core tenets of actionism, the paper is split into four sections assessing its theoretical struts. Each of the central claims will be analyzed separately and tested using what we know about the role of sniffing in olfactory perception.

3. Direct Perception – Actionism as Theory of Access

Actionism is at its heart a theory of access – how we access the sensory properties present in the world. Though it is not clearly said, Noë endorses a version of direct perception according to which the sensory properties present in the world are directly perceived through the deployment of sensorimotor skill based knowledge. The knowledge and implicit deployment of sensorimotor contingencies allows for the transition from sensation to perception. However, the sensory qualities are directly sensed without any need for further encoding or representational mechanisms.

> My proposal is that what brings the world into focus for perceptual consciousness is our understanding of the ways movement alters sensory events. Mere sensation does not rise to the level of perceptual experience for perceptual experience we need sensation that we understand. (Noë [2008], p. 532).

The tacit assumption of a theory of direct perception helps situate the theory and highlights that perceptual access is the focus of actionism. It is unlikely that olfactory perception will cause much trouble for this assumption of direct perception, as it is arguably the case that what we smell is the molecular structure of chemical compounds within odor plumes (Young [2011]). The olfactory quality of an olfactory object is determined by these chemical properties as they engage with the olfactory epithelium. The olfactory qualities are objective properties of the external world that we sense through our inhalation of these objects in our surroundings. Hence, a direct theory of perception seems most fitting for the nature of olfactory perception and in keeping with actionism. However, the further claim regarding sensorimotor contingencies being necessary and sufficient without further encoding or representational mechanisms needs testing.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. The Deployment of Sensorimotor Contingencies

The third claim that our perception of objects depends upon implicitly deploying sensorimotor contingencies faces little opposition from olfaction, but it requires clarification. The existence of law-like regularities between our movements and the way things appear is not enough. What needs to be established is that perception is only possible through the implicit deployment of this skillful knowledge. We need neither explicitly know these sensorimotor contingencies in a manner that is reportable in a propositional manner nor deploy them in an on-line routine like a set of rules or look up table. Rather the deployment of our sensorimotor knowledge is a skillful type of action performed in an automatic fashion. Noë's most explicit rendering of this implicit knowledge is:

> What matters for my purposes is that (i) perceivers are ***familiar*** with the way sensory motor stimulation varies as a function of movement; (ii) perceivers are generally unable to say what the relevant patterns are; (iii) being able to say what they are would not, in and of itself, be evidence of possession of the relevant perceptual capacities. (Noë [2006], p. 33 [emphasis added])

The last two claims are not problematic if olfactory perception is constituted by the sensorimotor contingencies, as these are mostly unattended and implicit. In general, we do not attend to our olfactory experiences, and are not conscious of most of the olfactory perceptible objects coming in contact with our sensory system. Based on the temporal processing time involved in sniffing it has been argued that olfaction is analogous to a constant state of change blindness (Sela and Sobel [2010]). The dearth of awareness to olfactory experience and our sniffing behavior is in keeping with the last two conditions, but the first is problematic.

I am dubious that we are *familiar* with how our movement affects our experience of smell. We might recognize that to locate a smell, we move around to gain access to the odorant's gradient. However, even this very limited form of movement dependent olfactory perception might escape peoples' grasp. Furthermore, we do not recognize that we modulate our sniff responses in a very robust and fine-grained manner to the valence of an olfactory object. Within 150 ms of the onset of a sniff the volume of air intake and strength of motoric inhalation are modulated in accordance with the pleasant or unpleasant nature of the stimulus (Johnson et al. [2003]). Odorants that are pleasing are inhaled more deeply and strongly, while those rated as unpleasant are sniffed less vigorously and we attempt not to inhale them (Bensafi et al. [2002, 2003]).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Doubtlessly the third tenet of actionism is applicable to olfactory perception based on the implicit role of sensorimotor contingencies in our perceiving the valence of an olfactory object. However, the extent to which (i) can be met depends upon the level of *familiarity* that is presupposed. Requiring too much olfactory familiarity will only breed contempt for actionism, due to humans' rarely attending to their active breathing patterns in relation to smelling. Furthermore, if too little familiarity is assumed then the theory is at best a theory of non-conscious sensory encoding and no longer perception. If (i) above requires some manner of conscious reportability of our olfactory experience and familiarity thereof then actionism will come up short; while if Noë allows for non-conscious processing this is in keeping with the experimental literature on the sniff sequence reviewed in section 2 except that the theory's target would be non-conscious states and not the phenomenological experience of perceptual states.

5. Perception Is Constituted by Motor-sensory Dependence

Actionism's most contentious claim is that the transition from sensation to perception is constituted by the subject's sensorimotor knowledge. The current section clarifies the role of sensorimotor knowledge in perception and explains why it is most likely not true of olfactory perception. According to actionism, we have direct access to objects in the environment, yet there needs to be an intermediate step that explains the transition from sensations to perception. The chief motivation of the project is an attempt to explain the perceptual presence of perceived objects in their entirety when we only sense surfaces or parts of objects. The transition from sensation to perception generates the necessity of sensorimotor contingency. "Sensory stimulation is intelligible only if its relation to us and to things around us is comprehensible" (Noë [2008], p. 535).

5.1. The Primacy of Motor-sensory Dependence

Intuitively our knowledge of how our interaction with our surroundings changes our sensations of the environment allows us to perceive objects in their entirety. However, the roles of the sensory and motor parts of sensory-motor knowledge within the theory need clarification. Are the sensory and motor components separable? If they

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

each perform independent functions, then which plays the primary role in generating perception?

Noë is explicit that the motoric component generates the changes in sensory stimulation. Even though it is often referred to as sensory-motor contingencies what is really meant is motor-sensory dependence. Carefully stated, the claim is that our knowledge of how our movements effect sensory stimulation allows us to experience the world as present and available for interaction.

> The detail shows up not as 'represented in my mind', but as available to me. It shows up as present – this is crucial – in that I understand, implicitly, practically, that by the merest movement of my eyes and head I can secure access to an element that now is obscured on the periphery of the visual field. It now shows up as present, but out of view, in so far as I understand that I am now related to it by familiar patterns of motor-sensory dependence. It is my basic understanding of the way my movements produce sensory change given my situation that makes it the case, now, even before I have moved an inch, that elements outside focus and attention can be perceptually present. (Noë [2009], p. 474)

The motoric component is not only primary in generating the transition from sensation to directly perceiving reality, it is necessary. "The obtaining of sensorimotor contingencies is necessary but not sufficient for perceptual consciousness" (2008, p.536-7). What fills out their further sufficiency is the implicit mastery of said contingencies. The implicit use of sensorimotor knowledge in olfactory perception is certainly within keeping of actionism, however, it will be argued that while sniffing is certainly sufficient for generating olfactory perception it is not necessary.

Moreover, actionism does not assert that the motor-sensory dependencies have a causal or determining relation in generating perception. Rather, Noë makes a stronger claim that perception is constituted by our knowledge and implicit deployment of motor-sensory contingencies. Logically this amounts to the conditional that if someone is undergoing a perceptual experience then they must have motor-sensory knowledge that is being masterfully deployed. Furthermore, the constitutive claim is equivalent to denying that one can have perception without motor-sensory dependence. Thus, what will be shown is that we can have olfactory perception in the absence of the motoric component of sniffing thereby yielding a contradiction between actionism's key claim and what is known about olfactory perception. To resolve the contradiction there are two possible conclusions; either the denial of perception being constituted by motor-sensory dependences, or that

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

olfactory experience is perceptual. The rest of this section will survey the literature on sniffing in favor of the first option, while section 6 will provide reasons to think that olfactory experiences should be considered perceptual.

5.2. The Motoric Component of Sniffing

Sniffing varies depending upon the concentration of the odorant plume, odorant intensity, and the presence or absence of an odorant. Nevertheless experiencing stimuli with olfactory qualities (i.e. what it smells like) does not require sniffing. The somatosensory experience of airflow and stimulating the olfactory epithelium are sufficient for the perception of smells, but sniffing and the motor component in particular are not necessary.

The necessity of airflow itself might be questioned based on a number of experiments whereby subjects had their nasal cavity flooded with an odorant-laced liquid to see if it elicited a sensation of olfactory quality. However, the reported results of these experiments vary from some claiming elicit olfactory experiences (Veress [1903]), to others who do not (Proetz [1941]; Weber [1847]) or do but with varying degrees (for a full discussion see Moncrieff [1946]). Furthermore, it might be questioned if these experiences might be attributed to olfactory perception, trigeminal stimulation, or somatosensory stimulation. To control for such issues Bocca (1965) delivered odorants to the olfactory epithelium by injecting them into blood circulation thereby delivering olfactory stimuli to the sensory transducers without sniffing or airflow. His results indicated that without active sniffing subjects do not report perceiving any olfactory qualities. So while odorant laced airflow might not be necessary, something about sniffing seems to be required.

One explanation of Bocca's results is that the delivery of odorants to the olfactory epithelium is not sufficient, as the mechanical stimulation of the epithelium is also required for producing the experience of smell. However, two other explanations are possible: the somatosensory experience of a medium flowing through the nostrils is required for the perception of smell; or, alternatively, the motoric action of sniffing and behaviorally modulating our nostrils is the necessary component in producing the smell experience.

Regarding the first option, in one of a series of experiments Sobel et al. (1998) demonstrated that subjects could perceive olfactory qualities even if the somatosensory experience of airflow was inhibited by topical anesthetic. Even in the absence of experiencing air flowing through our

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

nostrils, we nonetheless perceive olfactory qualities. However, this merely demonstrates the absence of the somatosensory stimulation does not affect our capacity to perceive olfactory qualities. To generate the contradiction with actionism it needs to be shown that perception occurs in the absence of the motoric component. To demonstrate this we must turn to the other set of experiments conducted by Sobel et al. (1998).

Aside from the aforementioned experiment using topical anesthetic, Sobel et al. (1998) conducted three experiments that fully clarify the role of the motoric component in sniffing. By fully occluding the nostrils, such that no air could flow through the nasal cavity when sniffing they showed that the motoric component alone does not generate the perception of smells. What is more interesting though is that by passively blowing air at the nostrils the somatosensory experience of airflow elicited activation in the relevant olfactory areas of the piriform cortex. Furthermore, even odorless air passively presented to the nostrils without sniffing elicits an experience of olfactory perception. To summarize, sniffing is not necessary for us to perceive smells and even when sniffing is used to gain access to the olfactory realm of stimuli the motoric component is inessential. Moreover, the somatosensory experience of airflow can be a sufficient condition of undergoing olfactory experiences – even passively.

> What would undercut it [the enactive approach] would be the existence of perception in the absence of bodily skills and sensorimotor knowledge which, on the enactive view, are constitutive of the ability to perceive. Could there be an entirely inactive, an inert perceiver? (Noë [2006], p. 9)

In answer to Noë's question, yes. In the case of olfactory experience we can have the perception of olfactory qualities completely passively. Thus, the central tenet of the theory that perception is constituted by the knowledge and mastery of motor-sensory dependences seems to be falsified when we consider active sniffing.

The experimental evidence demonstrates that the motoric component of sniffing is inessential for olfactory perception. Yet, the determinate role of the sensation of airflow in olfactory perception might provide some refuge for actionism. Being charitable it could be replied that the reason the airflow elicits the piriform activation is because usually our motor action of sniffing is what brings in and creates the airflow through the nostrils, thus we are implicitly deploying our knowledge of motor-sensory dependence.

Developmentally we slowly acquire mastery of this dependence, thus even in the absence of motoric action we deploy our implicit knowledge of how this sensation is generated in normal circumstances. In his earlier

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

work on enactivism, Noë emphasizes the importance of the developmental acquisition of sensorimotor knowledge. Yet, the developmental line of reply will not help in this instance, as the olfactory system is on-line and allows us to perceive the olfactory qualities of odorants as neonates if not *in utero* (Stein et al. [1958]; Steiner [1977]; Schmidt and Beauchamp [1988]; Schmidt [1992]).

Assuming by this point that the contradiction has been proven that we can have olfactory perception without the motoric component of motor-sensory dependence, Noë could retreat to the claim that olfaction is not a perceptual modality. But before we go down that road it might be wondered what role sniffing plays if, as is argued above, we can perceive smells without sniffing. Why do we sniff if not to perceive olfactory objects in the environment?

6. Perceiving Olfactory Objects – Perceptual Presence

The fourth tenet of actionism, which also serves as its *explanandum* is the phenomenal experience of perceptual presence. Though we do not receive sensory stimulation from the entirety of a three-dimensional object, we nonetheless perceive the object as a complete entity. According to actionism, the *explanans* of this phenomenon depends upon knowledge of perceptual object constancies. Our knowledge of motor-sensory dependencies facilitates filling in the sensory information, thereby generating the experiential percept of punctate entities external to us. Since Noë's theory is constructed to account for our phenomenological reports of consciously attended perceptual objects, the alternative conclusion provided by the previous section that olfaction is not perceptual might not seem so drastic.

We are not continually conscious of our olfactory experiences and rarely attend to them (Sela and Sobel [2010]). Furthermore, when we do attend to smells they seemingly appear within our nose in an almost unexpected manner (Peacocke [2008]). Moreover smells do not seem to present themselves to us as olfactory objects in the same manner as visual objects. Smells are not spatially or temporally punctate - they have vague boundaries that are difficult to individuate. Our experience of smells does not seem to present us with distal olfactory objects with a locatedness, rather we experience them as being somewhere in our surroundings (Batty [2010]).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

6.1. Perceiving Olfactory Objects

As it is a chemosense what we olfactory perceive is the molecular structure of chemical compounds within odor plumes, thereby making our olfactory experiences of olfactory quality (i.e. what something smells like) a perception of mind independent of qualitative objects within the environment. However, the nature of smells relative to the external object of perception need not go any further about how we claim to experience odors *as* olfactory objects. The experience *as* is the operative issue, since the matter at hand is how we conceive of the object of olfactory perception when we consciously attend to olfactory experiences.

Most can attest to their ability to recognize and identify the smell of freshly brewed coffee. Furthermore, we can identify this reoccurring smell against the background of other breakfast odors, such as the doughy smell of pancakes or the intoxicating smell of sizzling bacon. We do not smell a smudge, rather our experience presents us with a multitude of olfactory objects within an olfactory array, such that one might be so inclined to think these odors compose scenes (i.e. smellscapes).

Our olfactory experience of breakfast attests to our psychological ability to track the objects of olfactory experience through their temporal changes in features, as well as their combination and mixture with other olfactory objects. When attention is paid to this aspect of our olfactory experience it becomes apparent that these are objective experiences of mereologically complex entities. Considered in this light the objective status of the entities experienced using olfaction is supported by ecological theories of olfactory perception (Wilson and Stevenson [2006]; Gottfried [2010]). The methodological assumption of these theories is that the olfactory object is identified with the complex set of molecular compounds that we psychologically group together in tracking, locating, and securing objects that are of value to us in maintaining our homeostatic needs.

Wilson and Stevenson (2006) develop the most exhaustive scientific account of the object of olfactory perception in keeping with the criterion of a perceptual object as a mereologically-structured entity. To explain our psychological ability to parse the chemical sea in which we are immersed into temporally persisting recognizable objects, they provide a host of evidence that olfactory object perception partially depends upon synthetic processing. Their ecological theory supports the claim that we experience olfactory objects in a mereologically complex fashion, thereby establishing that olfactory experience and perception are object directed.

Additionally, our experience of odors satisfies the criterion of figure-ground segregation, which is instrumental in ascertaining the objective nature of perceptual entities that do not fulfill the rigid requirements of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

spatial locatedness. Empirical evidence for olfactory figure-ground segregation may be garnered from the overshadowing effect in odor mixture qualities. When combining odorants in a mixture, if the constituents smell similar on their own it is often difficult to recognize these constituents within the mixture; conversely, if they smell dissimilar on their own it is often quite easy to distinguish them within a mixture. However, in every variation of similar and dissimilar pairings of odorants there is "evidence of overshadowing of one component by another, depending upon the concentration level" (Kay et al. [2005], p. 727). Furthermore, if the concentration level of the overshadowed item is increased it is possible to switch the overshadowing effect. Indeed, whether one smells an odor *a* against a background of odor *b* (or vice-versa) can be manipulated by altering the concentration levels of the components of the mixture.

6.2. The Role of Sniffing in Olfactory Object Perception

Given that we do perceive smells as perceptual olfactory objects external to us and that we can have the experience of a smell independent of sniffing, why do we sniff? It is arguably the case that what determines the experience of olfactory quality is our perception of the molecular structure of chemical compounds within odor plumes. Such that the qualitative character of the olfactory experience is generated by the molecular structure of the chemical compound coming into contact and being transduced at the olfactory epithelium. However, a sufficient concentration level of odorants is required both for the experience of the olfactory quality and our capacity to perceive the smell as an objective entity external to us. What this suggests is that the olfactory plume plays a role in our ability to perceive smells as olfactory objects. So why then do we sniff?

Sniffing facilitates the actively exploration of a chemical smellscape in a manner that enables tracking the concentration and intensity of odors across a landscape. Sniffing enables the encoding of the intensity and concentration of an odorant (Sobel et al [1998]; Mainland and Sobel [2006]), thus it might help determine the spatiotemporal nature of olfactory objects. Odor plumes contain gradients of odorant concentration, yet we can recognize smells through changes in their intensity.

Shifts in concentration levels of an odorant generate variation in olfactory qualities. So how is it that we are able to have the perception of object constancies, such as our ability to recognize the smell of coffee across multiple exposures of varying concentrations? Recent experimental

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

research has shown that naïve mice treat odor plumes of varying concentration and ratios of the same kinds of chemical components as being of different qualities (Cleland et al. [2012]). Future developmental research on naïve humans is required to corroborate that odor plumes composed of the same chemical compounds at varying concentration levels are perceived as having different qualities. But, assuming this effect is not species specific, these results indicate that some properties of the odor plume partially determine olfactory quality.

Our ability to recognize an odor as having the same smell across presentations of varying concentrations is an acquired capacity determined by concentration invariance, which depends upon learnt odor categorization. Concentration invariance extends beyond the perceptible properties presented by the external object of olfactory perception including its property of olfactory quality. Further research needs to be done on Humans' olfactory perceptual constancy of concentration invariance. Speculatively, it is determined in accordance with the ratio of the chemical compounds within a given odor mixture, since the compositional ratio of components should stay constant despite a shift in concentration levels (Uchida et al. [2007]). Acquiring the capacity of concentration invariance depends upon multiple exposures to the same chemical compounds across varying concentrations, which is in keeping with the suggestion that active sniffing facilitates tracking the variations in concentration and intensity across exposures thereby enabling perceived object constancies of distal odor stimuli.

Doubtlessly our ability to parse the chemical sea of smells on a daily basis requires sampling across time and movement, as well as acquired knowledge of concentration invariance based on the ratios of constituents composing an odor. However, our perception of mereologically complex smells across time need not depend upon our knowledge (implicit or otherwise) of olfactory motor-sensory contingencies, as it might depend upon tracking the somatosensory experience of airflow through the nostrils.

We can parse our chemical environment into individual odors within an unfolding smellscape, which speaks to olfaction being a perceptual modality. If sniffing produces our capacity to demarcate the boundaries of smells thereby individuating and identifying them, then sniffing is required for olfactory object perception and not mere sensations of olfactory quality. Having established that olfactory experiences are perceptual, we can now turn to testing the fourth tenet of actionism, that we perceive objects as complete within an external environment.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

6.3. Locating Smells and Individuation of the Olfactory Modality

We locate smells using differences in concentration, however, this requires active exploration either through movement of the entire body or at the very least our head (Richardson [2011]). The difference between the odorant properties presented to each nostril enables our ability to track an olfactory stimulus through an environment over time (Porter et al. [2007]). However, if we consider our olfactory experiences as individuated just in terms of olfactory stimulation as transduced only at the olfactory epithelium then smells are not localizable at a given time.

Despite evidence that each nostril creates a different olfactory percept as demonstrated by binaural rivalry between the nostrils (Zhou and Chen [2009]), it has been shown that we cannot tell at a given instant if an odorant is present in the left or right nostril (Radil and Wysocki [1998]; Frasnelli et al. [2008]). Moreover, actively sniffing a pure olfactory odorant (i.e. a stimulus that does not stimulate the trigeminal nerve endings within the nose) does not allow us to localize the odorant (Frasnelli et al. [2009]). Olfactory perception can, across time (diachronically), have spatial structure, yet at any particular time (synchronically), olfactory experience has no spatial structure even when active sniffing is taken into account.

Instances of perception that present us synchronically with a localizable smell or an olfactory object with a distal location are mediated by trigeminal stimulation within the nostrils. Trigeminal stimulation allows subjects to distinguish the onsets of stimulation between nostrils (Kleeman et al. [2009]), as well as to localize odorants within 7-10 degrees of their location (von Bekesy [1964]). Furthermore, if an odorant contains chemicals that elicit trigeminal activation then active sniffing further enables us to localize the distal extent of the olfactory object (Franselli et al. [2009]), thus leaving us with a conundrum that might help clarify actionism, as well as an issue concerning how to individuate the senses. Olfactory quality (what something smells like) is determined based on transduction of odorants at the olfactory epithelium independent of trigeminal stimulation. Trigeminal stimulation allows us to localize gaseous clouds that constitute odors, but it does not generate the perception of olfactory qualities independent of stimulation of the olfactory epithelium, thus creating a dilemma for actionism regarding how to individuate the olfactory modality.

The first option is to individuate olfaction as a perceptual modality including trigeminal stimulation. Considering olfaction as everything going on inside the nostrils gives us instances of perceived olfactory

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

objects with spatiotemporal properties within our surroundings. Furthermore, this allows for an active role of sniffing in olfactory perception of odor objects as being locatable and perceptually present to us. However, this option then falsifies actionism because the motoric component is neither necessary nor constitutive of olfactory perception. Moreover, the somatosensory component is necessary for our olfactory perception in a manner that allows for inert olfactory perception.

The second option is to individuate olfaction based on olfactory qualities, such that the only time we should say we perceive olfactorily is when we are presented with the experience of a smell. Excluding the other sensations from within the nostrils from the olfactory modality provides a retreat for actionism. It could be replied on their behalf that the experience of smelling without active sniffing do not yield synchronically perceived olfactory objects within the environment. Therefore, given the fourth tenet of actionism we can never make the transition from mere sensations to perceptions.

Olfactory experience considered synchronically might allow for the alternative conclusion in section 5 that the olfactory sensory modality never rises to the level of being perceptual. Yet, we can fill in the spatial aspects of an olfactory object even under this construal of the olfactory modality using diachronic active exploration making the restriction of its fourth tenet to synchronic perception the only bastion of hope for actionism.

A possible and overly charitable reply on their behalf could be that when considering our ability to perceive objects using perceptual constancies one of the necessary components is the felt presence of the object as having a locatedness in three dimensional space relative to us. Perceived locatedness is a stronger requirement whereby at a given instant we perceive the object as being at a given place within a spatial array. Immediately upon opening our eyes we are seemingly presented with a spatial array of visual objects located before us. Phenomenologically this experience is unmediated by our movement in a way that highlights what might be the key difference between olfaction and visual. Vision presents us with object locatedness at an instance, but olfaction does not, thus synchronic olfactory experiences are not object directed and perceptual in the proper sense specified by the theory. Our perceptual experience considered from a fully conscious introspectively reportable state seems to bear out this claimed difference, and provides a possible retreat, but there are two possible replies. The first questions the veracity of claimed lack of movement in the visual experience, and the second relies on a thought

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

experiment to account for the difference in size between perceptible punctate visual objects and diffused gaseous olfactory objects.

6.4. Diachronic Olfactory Perception

The assumption that olfaction automatically presents us, synchronically, with spatially located objects can be challenged on the grounds that similar diachronic processes occur in vision. To see things, one's eyes must be in constant motion either through volitional control or through saccadic and micro-saccadic movements. If one's eyes were to stop moving the visual field would shrink and eventually turn a uniform grey.[1] Saccadic eye movement presents a *prima facie* analogy to the role of active diachronic exploration through movement or sniffing in olfaction. If saccadic eye movement is not excluded either as a part of the perceptual experience or as an enabling condition of the visual percept, then there seems to be no reason to exclude exploratory olfactory movements.

Anticipating this reply, it could be argued that olfactory active exploration is not equivalent to saccadic eye movements. While the latter is not under volitional control, the former is always required when locating and moving towards a smell. However, some forms of saccadic eye movement required for visual perception are under our control in an analogous way. Thus, a more charitably interpretation of actionism is that it concerns the phenomenal awareness of movement such that one is not usually aware of saccadic eye movements, but always aware of movements used in attempting to locate a smell; raising the question of how our phenomenological awareness presents smell's temporal aspect in a manner unlike the spatiotemporal bound entities of vision.

Given the motivation of actionism as a departure from employing vision as our theoretical launchpad there would be a perverse irony in holding olfaction accountable to the phenomenology of visually presented objects. Undoubtedly there are phenomenological differences between the perceptual modalities, but further argumentation is needed before we can adjudicate the dominance of perceptual objecthood criteria between the modalities. As such I ask the reader to consider the following thought experiment using olfaction as our starting point.

[1] For an at home demonstration of this phenomena simply immobilize your eyeballs by holding them firmly against the side of the eye socket.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

6.5. Perspectival Shrinkage

Assuming that the olfactory object of perception is chemical structures within gaseous odor plumes, then these are large diffuse entities that cannot be fully perceived in a single olfactory scene. Demarcating the edges of an odor requires either multiple samples of the stimulus to determine its concentration gradient or movement relative to the edges of the object. Now consider visual objects: these are experienced as three-dimensional punctate entities presented to use with a determinable location. But what if we were to shrink down to the size that we could move within these visual objects. Presumably we could still perceive them visually – they would still have refractory properties that our visual system is sensitive to. But in order to perceive their surfaces and edges we would need to move about.

In this situation my intuition is that we would still see visual objects, but that they would be without their phenomenological locatedness synchronically perceived. Holding the proper sensible constant, while shifting our perceptual perspective through shrinkage allows one of two possibilities: we conclude that vision is not perceptual in this scenario; or vision is still perceptual, but with further diachronic exploratory movements built in to allow for objectual perception.

Unsurprisingly, it is my contention that actionism must allow for diachronic exploration as part of the perceptual process, thus even on the second manner of individuating the olfactory modality we can perceive objects. Though this yields an unpleasant conclusion for the theory, it is in keeping with its overarching claim that perception is generated by both our knowledge of sensorimotor contingencies and our deployment of them. The down side of this conclusion for actionism is that even without sniffing and trigeminal stimulation we could undergo olfactory perception by tracking the somatosensory component of airflow through the nostrils independent of the motoric component of movement. Nevertheless it could be replied that even in these instances some manner of implicit knowledge is being deployed based on past movements, which calls for future experimental studies that argue olfaction is best suited for in isolating the motoric and sensory components necessary for our perception of smells.

7. Conclusion

Sniffing plays an important role in our experience of smells, but it does not substantiate actionism's claims regarding the necessity of our knowledge and implicit masterful deployment of motor-sensory dependencies. While

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the theory is fitting given its foundations as a theory of direct perception, as well as the claim that our experience of smell is implicitly generated by our use of actively sampling and modulating our sniffing behavior, the further claim that motor-sensory dependencies constitute the transition from sensation to perception is not supported by the experimental literature surveyed. By clarifying the exact nature of actionism's claim regarding sensorimotor contingencies it was noted that the motoric component was the decisive factory in determining perception, and that sniffing served as a fitting test case since the motoric and somatosensory components could be isolated.

Not only can we have olfactory perception passively, but even when sniffing is employed in generating an olfactory percept the motoric component is inessential. Perhaps the motor stimulation is sometimes causally required, but the stronger constitutive claim is in no way substantiated. To handle the possible conclusion that olfaction is not perceptual, it was further shown that our olfactory experience is of olfactory objects in the environment in a manner that satisfies even the fourth tenant of actionism. We perceive olfactory objects in the environment in a manner that allows us to recognize and track them through spatiotemporal changes, as well as shifts in concentration and intensity. Olfaction is not only an applicable test case, but also a good fit for testing the key tenets of actionism. Nonetheless it seems as if this form of enactivism has taken its last breath, yet smelling endures.

References

Arzi, A., Shedlesky, L., Secundo, L. & Sobel, N. (2014), Mirror sniffing: humans mimic olfactory sampling behavior. Chemical senses 39 (4): pp. 277-281.

Batty, C. E. (2010), What the nose doesn't know, J. Conscious, Stud. 17, pp. 10–17.

Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2002), Autonomic nervous system responses to odours: the role of pleasantness and arousal, Chem. Senses 27, pp. 703–709. DOI: 10.1093/chemse/27.8.703.

Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2003), Perceptual, affective, and cognitive judgements of odors, Brain Cogn. 31, pp. 270–275. DOI: 10.1016/S0278-2626(03)00019-8.

Bocca, E., Antonelli, A. R. & Mosciaro, O. (1965), Mechanical co-factors in olfactory stimulation, Acta Oto-laryngol., 59, pp. 243–247.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Cleland, TA, Chen, ST, Hozer, KW, Ukatu, HN, Wong, K. J. & Zheng, F. (2012), Sequential mechanisms underlying concentration invariance in biological olfaction. Front. Neuroeng. **4**:21. DOI: 10.3389/fneng.2011.00021.

Galef, B. B. (2013), Animal Communication: Sniffing Is About More Than Just Smell, Current Biology, Volume 23, Issue 7, 8 April 2013, Pages R272-R273.

Gottfried, J. A. (2010), Central mechanisms of odour object perception, Nature Reviews Neuroscience 11, pp. 628–641.

Gray, R. & Tanesini, A. (2010), Perception and Action: the taste test, Philosophical Quarterly, 60:241, pp. 718–73.

Frasnelli, J., Charbonneau, G., Collignon, O. & Lepore, F. (2008), Odor localization and sniffing, Chem. Senses 34, pp. 139–144. DOI: 10.1093/chemse/bjn068.

Kay, L. M., Crk, T. & Thorngate, J. (2005), A Redefinition of Odor Mixture Quality, Behavioral Neuroscience, 119(3): pp. 726-733.

Kleemann, A. M., Albrecht, J., Schöpf, V., Haegler, K., Kopietz, R., Hempel, J. M., et al. (2009), Trigeminal perception is necessary to localize odors, Physiol. Behav. 97, pp. 401–405. DOI: 10.1016/j.physbeh.2009. 03.013.

Mainland, J.D. & Sobel, N. (2006), The sniff is part of the olfactory percept. Chemical Senses 31 (2): pp. 181-196.

Moncrieff, R. (1946), The Chemical Senses, New York: John Wiley & Sons.

Noë, A. (2001), Experience and the active mind, Synthese 29, pp. 41-60.
Noë, A. (2001b), Experience and experiment in art, Journal of Consciousness Studies, 7.

—. (2002), On what we see, Pacific Philosophical Quarterly, 83: 1: pp. 57-80.

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

—. (2005), Real presence. Philosophical Topics, 33(1), pp. 235–264.

—. (2006), Précis of action in perception**,** Psyche, 12:1.

—. (2007), Understanding Action in Perception: Replies to Hickerson and Keijzer, Philosophical Psychology, 20:4, pp. 531-538.

—. (2009), Conscious Reference, Philosophical Quarterly, 59:236, pp. 470-482.

O'Regan, J.K. (1992), Solving the "real" mysteries of visual perception: the world as "outside memory", Canadian Journal of Philosophy 46, no. 3: pp. 461-488.

O'Regan, J.K. & Noë A. (2001a), A sensorimotor approach to vision and visual consciousness, Behavioral and Brain Sciences 24, 5: XXX.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

O'Regan, J. K. & Noë A. (2002), What it is like to see: a sensorimotor theory of perceptual experience, Synthese: 29: pp. 79-103.

Olofsson, J. K. (2014), Time to smell: a cascade model of human olfactory perception based on response-time (RT) measurement, Front. Psychol., 5:33. doi: 10.3389/fpsyg.2014.00033.

Lycan, W. G. (2006), Enactive Intentionality, Psyche, 12:3.

Peacocke, C. (2008), Sensational properties: theses to accept and theses to reject, Rev. Int. Philos. 62, pp.7–24.

Porter, J., Craven, B., Khan, R. M., Chang, S. J., Kang, I., Judkewitz, B., et al. (2007), Mechanisms of scent-tracking in humans, Nat. Neurosci, 10, pp. 27–29. DOI: 10.1038/nn1819.

Prinz, J. (2006), Putting the brakes on enactive perception, Psyche, 12:3.

Proetz, A.W. (1941), *Applied Physiology of the Nose*, St Louis, MO: Annals Publishing Co.

Radil, T. & Wysocki, C. J. (1998), Spatiotemporal masking in pure olfaction, Ann. N. Y. Acad. Sci. 855, pp. 641–644. DOI: 10.1111/j.1749-6632.1998.tb 10638.x.

Richardson, L. (2011), Sniffing and smelling, Phil. Stud. 162, pp. 401–419. DOI: 10.1007/s11098-011-9774-6.

Sela, L. & Sobel, N. (2010), Human olfaction: a constant state of change-blindness, Experimental Brain Research 205(1): pp. 13-29.

Sobel, N., Prabhakaran, V., Desmond, J. E., Glover, G. H., Sullivan, E. V., Goode, R.L. & Gabrieli, J. D. E. (1998), Sniffing and smelling: separate subsystems in the human olfactory cortex. Nature 392 (6673): pp. 282-286.

Sobel, N., Khan, R. M., Sullivan, E. V. & Gabrieli, J. D. E. (2000), The world smells different to each nostril. Nature 402 (6757): 35.

Schmidt, H. J. (1992), Olfactory hedonics in infants and young children, in Fragrance: The Psychology and Biology of Perfume, eds. S. Van Toller and G. H. Dodd, London: Elsevier, pp. 195–219.

Schmidt H. J. & Beauchamp G. K. (1988), Adult-like odor preferences and aversion in three-year-old children. Child Dev. 59, pp. 1136–1143. DOI: 10.2307/1130280.

Stein, M., Ottenberg, P. & Roulet, N. (1958), A study of the development of olfactory preferences, AMA Arch. Neurol. Psychiatry 80, pp. 264–266. DOI: 10.1001/archneurpsyc.1958.02340080134028.

Steiner, J. E. (1977), Facial expressions of the neonate infant indicating the hedonics of food-related stimuli, in Taste and Development. The Genesis of Sweet Preference, ed. J. M. Weiffenbach (Bethesda, MD: NIH-DHEW), pp. 173–188.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Uchida N. & Mainen Z. F. (2007), Odor concentration invariance by chemical ratio coding, Front. Syst. Neurosci. 1:3. DOI: 10.3389/neuro.06.003.2007.

Veress, E. (1903), The irritation of the olfactory organs through the direct impact of odorous liquids, Pflugers Arch Gesamte Physiol Menschen Tiere, 95, pp. 365–408.

von Bekesy, G. (1964), Olfactory analogue to directional hearing, J. Appl. Physiol. 19, pp. 369–373.

Weber, E. H. (1847), Ueber den Einfluss der Erwärmung und Erkältung der Nerven auf ihr Leistungsvermögen, Arch. Anat. Physiol wiss Med Berlin, pp. 342–356.

Wesson, D. W. (2013), Sniffing behavior communicates social hierarchy, Curr. Biol., 23 (2013), pp. 575–580.

Wilson, D. A. & Stevenson, R. J. (2006), *Learning to Smell*, Baltimore, MD: The Johns Hopkins University Press.

Young, B. D. (2011), *Olfaction: Smelling the Content of Consciousness*, Doctoral Dissertation, City University of New York, The Graduate Center.

Zhou, W. & Chen, D. (2009), Binaural rivalry between the nostrils and in the cortex, Curr. Biol. 19, pp. 1561–1565. doi: 10.1016/j.cub.2009.07.052.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER FIVE

THE ONTOLOGY OF SOME AFTERIMAGES

BRYAN FRANCES

A good portion of the work in the ontology of color focuses on color properties, trying to figure out how they are related to more straightforwardly physical properties. Another focus is realism: are ordinary material objects such as pumpkins really colored? A third emphasis is the nature of what is referred to by the terms 'what it's like' or 'phenomenal character', as applied to color. In contrast, this essay is exclusively about select color tokens. I will be arguing that whether or not ordinary objects such as pumpkins are colored, regardless of what the true theory of color properties is, and independently of any talk of phenomenal character or what-it's-likeness, some afterimage experiences are very hard to fit into any plausible ontology, physicalist or not.

The Four Options for Afterimages

Suppose you have a very bright light bulb in the shape of a cow. The light is on and you stare at it from two feet away for about 30 seconds straight. Then you close your eyes tightly and put your hands over your eyes. What happens next is that you experience an afterimage, a cow-shaped blob of color that changes color over time. For me, it starts out orange, then it's red, then it's pink, then it's violet, and then it fades away. And it is roughly cow-shaped the whole time. Let's introduce some times into the scenario:

4:00:00pm: You close your eyes after staring at the brightly-lit cow-shaped light bulb.

4:00:01pm: You start experiencing a cow-shaped-and-orange-but-not-red patch.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

4:00:01pm: Someone breaks the light bulb; your eyes are still closed and you're still experiencing the cow-shaped-and-orange-but-not-red patch.

4:00:10pm: With your eyes still closed you gradually stop experiencing a cow-shaped-and-orange-but-not-red patch and start experiencing a cow-shaped-and-red-but-not-orange patch.

One truly has to go through something akin to this process in order to philosophize at all fruitfully about afterimages. At the very least, one should do it with an ordinary light bulb before reading much further.

When one has done so with the cow-shaped bulb, it will certainly *seem* as though there was this thing, *a cow-shaped-and-orange-but-not-red patch*, in your visual field from 4:00:01 to 4:00:10. Thus, it seems as though this is a correct description of how you started out:

There was a spread (expanse, blob, patch[1]) of color, call it X, that
(a) you were visually experiencing from 4:00:01 to at least 4:00:10 even though your eyes were closed,
(b) was cow-shaped (at least approximately, like a drawing of a cow) during that whole time,
(c) was orange during that whole time,
(d) was never red during any of that time.

I will use 'afterimage' as a term for X. I will use 'afterimage experience' for the visual experience, whatever it really is, that happens when one does the light bulb experiment. Our question in this essay is this: *what is X?* In other words, what is the thing that satisfies all four of (a)-(d)? I'll go over each of the only possible candidate answers:

1. X doesn't exist.
2. X exists and is an external physical thing (the surface of the light bulb perhaps).
3. X exists and is a non-physical thing.
4. X exists and is an internal physical thing (some part of the eye, eyelid, or brain perhaps).

[1] I use 'patch' in a neutral way so that the "patch" might be an ordinary external physical object (e.g., a light bulb), some part of the eye or rest of the human body, a part of one's non-physical soul, etc.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the pages that follow there are mistakes in my arguments. There *have* to be: I am going to be arguing against all the possible options for afterimages. My goal is to merely present the views and their main serious problems. My thesis regarding X is not that such-and-such an option is false but that each option faces difficult objections. My own view is that (4) is our best bet, but my arguments will not show even that relatively modest thesis.

Option 1: the Afterimage Does not Exist

Until recently no philosopher who worked extensively on perception—H. H. Price, Husserl, C. D. Broad, Ayer, Russell, Moore, and others—would even seriously consider denying that X exists. Nonetheless, many contemporary philosophers of perception hold that there is nothing that satisfies all of (a)-(d). At first, this may seem crazy: when one is having the afterimage experience it sure seems as certain as anything ever gets that there is a colored thing there!

1. If the afterimage X didn't exist, then when you closed your eyes you would experience just *darkness*, uninterrupted by any colored expanse. That's precisely what it would be like if there were no afterimages at all.
2. But of course that's not what happened when you closed your eyes. On the contrary, when you closed your eyes it wasn't just darkness! All one has to do is briefly look, and one will see that breaking the darkness was *color* and *shape* instantiated in something. The something in question may be subjective or ephemeral in various ways, but there was definitely *something there* that had color and shape.
3. Thus, X really does exist. So option (1) is false.

That is a reasonable argument, but the defenders of the 'X doesn't really exist' view have at least the beginning of an intelligent response: what exists is not an afterimage, which is a colored patch or spread or whatever you would like to call it, but an *experience of* the old light bulb, which is a kind of mental process or event you're having and which is not colored or cow-shaped. For comparison, when you see a star at night, sometimes what you are seeing no longer exists: the star blew up millions of years ago but the light from it takes so long to get to earth that it's still arriving here. So even though you are, at midnight, visually experiencing star X, X doesn't exist at midnight while you are experiencing it.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

More precisely, some advocates of the 'X doesn't really exist' option (1) offer the following analogy:

The Star Case	**The Light Bulb Case**
At time T1 the star existed and was giving off light	At time T1 the light bulb existed and was giving off light
At later time T2 it stopped existing (the star blew up)	At later time T2 it stopped existing (the light bulb was destroyed)
At even later time T3 we visually experienced the star	At even later time T3 we visually experienced the light bulb (via the afterimage experience)

Unfortunately, there are serious problems with *this* defense of option (1) (other defenses, below, are superior to this one). For one thing, it seems to most people that after you closed your eyes you were indirectly experiencing the light bulb *via the presence of the afterimage X*. That is, it's partly *in virtue of* the existence of the image that you are experiencing the light bulb: the presence of the image, at 4:00:02, was the primary (but not sole) *means* to experience the non-existent light bulb at 4:00:02. Here is the key point: the bulk of the reason you can, right now at 4:00:02 after the light bulb has been destroyed, experience the light bulb, is this: right now, at 4:00:02, there is an image that is in your visual field, and the image was generated in the appropriate way from looking at the light bulb. So, the image has to exist, otherwise you wouldn't be experiencing the light bulb after it was destroyed.

When it comes to the star case, the realist about afterimage X wants to know what it is *about* one's midnight visual experience that allows one to see the nonexistent star. It's all well and good to say that the current light coming into the eye somehow allows one to experience the nonexistent star, but what is it that the light is doing to make that happen? The realist has an answer: the light is helping to produce an image, a white one, that exists at midnight. The eliminativist about the white spot that exists at midnight has to come up with a story that doesn't require any midnight image at all.

This suggests that the star case is just as puzzling as the afterimage case. In the afterimage case we struggle to find out what X is; in the star case we struggle to discover what the midnight white spot is. So even though the star-light bulb analogy is not bad, all this shows is that the star

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

case is roughly as paradoxical as the afterimage one. So the analogy doesn't help option (1).

The option (1) advocates, who insist that X doesn't really exist, don't give up at this point. The most common reaction goes something like this:

> When you have an afterimage there is some process going on in your visual system that is *importantly similar* to what goes on in your visual system when you see an external object with your eyes open in normal circumstances (for the afterimage experience in question, it's a bit as though one is in a dark room and there is a blurry object before one's eyes). For instance, suppose you take some LSD and hallucinate a pumpkin. So you are, in some sense of 'experience', experiencing an orange patch in your visual field. But strictly speaking, there is no orange patch there; it's just a hallucination after all. Your eyes and brain are functioning *as though* you are seeing a real pumpkin, at least approximately, but that doesn't mean there is any actually existing orange patch. In the case of some afterimage experiences, what is going on in your visual system is relevantly similar to what goes on when you see something like a rainbow, hologram, beam of colored light, or the sky; but there is no real thing in one's afterimage experience.

So their idea is this, temporarily focusing on hallucinatory images instead of afterimages:

i. When you're hallucinating an orange pumpkin, you are having a certain visual hallucinatory experience E_{hall}.
ii. When you see a real orange pumpkin, in perfectly ordinary circumstances, you are having a certain ordinary visual experience E_{ord}.
iii. E_{hall} and E_{ord} are highly similar—similar enough that they seem pretty much the same visually from the inside, from the point of view of the one having the experiences (this is consistent with them being dissimilar in philosophically interesting ways).
iv. Even so, whereas in the case of E_{ord} there is an existent orange object (the pumpkin), *in the case of E_{hall} there isn't any orange object at all*.

Although (iii) can be intelligently rejected, the main questionable claim is the second part of (iv)—the claim that in the case of E_{hall} there is no orange object at all.

The problem with that part of (iv) is revealed when we try to figure out exactly *how* E_{hall} and E_{ord} are similar. On the face of it, the answer to what I will call *the challenge*,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Precisely *how* are visual experiences E_{hall} and E_{ord} so similar? In what *respects* are they so similar? What is it about them that *makes* them similar?

Is this:

> They are similar in the sense that for each one there is something that is orange, not red, and pumpkin-shaped: with one experience, E_{ord}, the thing in question is a pumpkin; with the other experience, E_{hall}, the thing in question is an image that looks like a pumpkin. It's the *presence* of the two orange-but-not-red-pumpkin-shaped things that *makes* the two experiences E_{hall} and E_{ord} so similar. The similarity is obvious and open to introspection. It is the presence of the orange image in E_{hall} that *makes* that experience so similar to E_{ord}; the experiences are alike *in virtue of* the fact that the image and the pumpkin are so similar. Thus, in the case of E_{hall} there is an orange-but-not-red-pumpkin-shaped thing, contrary to (iv).

The advocates of the 'X doesn't really exist' view have the burden of offering a detailed proposal on exactly what goes on—what has to exist—when one is experiencing the afterimage. Vague talk about how the hallucinating person is having a "visual experience" that is similar to the one a person has upon seeing a real pumpkin clearly won't do, as it fails to tell us *how precisely* the experiences are similar—and when we try to say what it is that makes the experiences similar, it is highly natural to end up admitting that X really exists.

The 'X doesn't really exist' people could try to say that the two experiences E_{hall} and E_{ord} are similar in that they share some key *representational properties* (Block [1983], Dretske [1995], Harman [1990], Tye [1995, 1997, 2000] set the stage for recent discussion). More to the point, the reason the two experiences seem the same from the inside, especially when it comes to color and shape, is nothing over and above the fact that they share those representational properties. Both experiences represent orangeness and pumpkin-shape, and that's the core of the reason they are similar. That's how we answer the challenge posed above.

Surprisingly, the critic of (1) can accept all the claims of the previous paragraph. There's nothing there that helps option (1)!

The problem is that even if this representationalist proposal is true, there's nothing in it that avoids the image. The reason is this: it seems as though part of what *makes* the hallucinatory experience E_{hall} represent orangeness, for instance, what is doing most (but not all) of the representing work here, is the presence of the orange image that looks like a pumpkin. That is, it's partly (not solely) *because* the hallucinatory experience involves an image that instantiates orangeness in a pumpkin-

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

drawing fashion that it "represents the world" as containing an orange pumpkin. In addition, it seems clear that the whole reason the person having E_{hall} *consciously thinks* about the color orange at all (assuming this is a case in which she does so think) is that she is aware of an orange patch: it's the presence of the patch (or blob, or spread, or whatever term you like to use) that got her thinking of orange in the first place. The approximate truth is the first claim, not the second:

> Part of what makes it the case that she represents orangeness is the presence of an orange patch.
>
> Part of what makes it the case that she experiences an orange patch is that she represents orangeness.

Hence, the critic of (1) can *accept* the thesis that what makes E_{hall} and E_{ord} so similar is completely exhausted by facts about representation or concepts or intentionality or the like. The thesis doesn't address the ontological issue.

However, there is a way to use the representationalist idea to support (1). In order to pull it off, the advocate of option (1) needs two claims, one negative and one positive, and it's the negative one that is pivotal:

> The *Positive* Claim: what makes E_{hall} and E_{ord} so similar from the inside when it comes to color and/or shape is completely exhausted by facts about representation.
>
> The *Negative* Claim: what makes E_{hall} and E_{ord} so similar from the inside when it comes to color and/or shape does not involve E_{hall} having any existent orange, pumpkin-looking thing.

The conjunction of the two claims defines the *representation-with-no-X* view, which is what the supporter of (1) is looking for. As pointed out above, the critic of (1) need not object to the positive claim (she need not accept it either). All that really matters is the negative claim, and I know of no good reason to accept it. Indeed, the literature rarely distinguishes the claims, and it is even rarer to encounter anything more than the mere assertion of the negative claim.

I think it's safe to say that the representation-with-no-X view, applied to the afterimage experience we started out with (so we are setting aside hallucinations now), will strike a person as counterintuitive *provided* two conditions hold:

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

a) It includes the negative claim that the afterimage experience involves nothing orange or cow-shaped (this is the analogue of the negative claim for hallucinations).
b) The person in question actually tokens the afterimage experience type a few times and attends to it.

When you attend to the colorful afterimage experience are you looking at something that only suggests or calls up or is *about* orangeness? Or is the orangeness *spread out on something*? To most people who perform the experiment—and we should not restrict ourselves to philosophers who write about color—the afterimage experience does not appear to *merely* indicate the color orange in something like the way thoughts, words, concepts, or imagined images do (I'll comment on the latter below). Instead, the orangeness is spread out right on the afterimage itself. One finds oneself looking at a cow-shaped spread of orangeness, and the idea that the spread represents orangeness *without including anything really orange* is observationally implausible. For most people, there is no need for an argument for the claim that the orangeness is actually instantiated in the afterimage experience. Instead, people tend to think that all one needs to do in order to see that orangeness is actually instantiated is just *look*. If a philosopher does the afterimage experiment, focuses her attention on the cow-shaped orange patch in the darkness, and still feels content to deny that anything at all in her experience is orange—so she is saying that there is nothing orange there—then she can pretty much convince herself of anything.

This is *not* an appeal to "philosophical intuition", in any of the senses of that phrase. Instead, it's an appeal to simple empirical *observation*. The observation in question has to be tempered somewhat, because as soon as one starts thinking hard about color one realizes that orange afterimage-expanses and orange pumpkins may not be orange in the same way (assuming pumpkins are orange; more on this issue below). My point here is that the advocate of the existence of X is not moved by any *argument*, such as the argument from illusion or hallucination. She is moved by her simple *observation* of a cow-shaped spread of orange in her experience: the idea that there is no spread of orange there is observationally refuted. It's not the case that the realist about X is arguing along the lines of 'If it visually appears that x is F, then something (perhaps not x) is F' (more on that principle below).

The advocate of (1) could say that what makes the afterimage experience represent orangeness and cow-shapedness is some complicated story, such as something akin to Fodor's about causal processes (Fodor

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

[1990]). The critics of (1) can accept that representation has a great deal to do with causal and/or other facts not open to introspection, but they will insist that the representationalists are ignoring the elephant in the room: *part* of the reason the afterimage experience makes one think of, or (what is different) otherwise represent, orangeness and cow-shapedness is the glaringly obvious one: when one is having the experience there is a really existing orange and cow-shaped thing that one is experiencing. This is part of the correct answer to 'How does it come to be that the experience represents orangeness and cow-shapedness?.[2]

Ian Phillips (2012) views the matter differently. He thinks that according to realists about X "we cannot adequately characterise experience solely in terms of a subject's apparent perspective on external, public reality"; he thinks the afterimage realist holds that afterimages "cannot be accounted for solely in terms of the ways in which apparent aspects of that world are presented to us" (2012, xxx). I think Phillips has mischaracterized the matter. The realist about X can admit that the representational aspects of the afterimage experience in some sense "exhaust" its philosophical interest. Perhaps the qualia property instances reduce, in some ontologically robust sense of 'reduce', to parts of representational property instances. All she has to do in order to admit these theses is to claim that the truthmaker for the representational story will include, as a part, X. Just because we can "adequately characterize experience solely in terms of a subject's apparent perspective on external, public reality" doesn't mean that that perspective does not include, as a part, the image X.

I am going to continue to present the arguments against (1) in a moment, but it's worth noting here that, truth be told, the *real* argument for option (1), the line of reasoning that actually *moves* philosophers, has nothing to do with representation and goes as follows. First, the advocates of (1) think that options (2)-(4), to be examined below, are *highly* implausible.

2. X exists and is an external physical object: the surface of the light bulb perhaps.
3. X exists and is a non-physical object.

[2] Similar arguments show the inadequacy of the defense of (1) that runs 'But the afterimage experience is an illusion; so, there is hardly any good reason to think there is anything orange there'. See also the remarks on Harman and Block below.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. X exists and is an internal physical thing: some part of the eye or brain perhaps.

Very briefly, they think (2) can't be right because the light bulb doesn't exist when the afterimage X does (on the assumption that X exists at all); they think (4) can't be right because no part of the eye or brain is orange and cow-shaped; hence, they think that realism about the afterimage (i.e., the denial of (1)) leads to the worst kind of dualism, view (3). But they also know that (1)-(4) are all the options; so they then grudgingly conclude that option (1) is the *least bad* view to take, since (2) and (4) are out and (3) is a disaster. Once a person has felt forced to settle for (1), the search for arguments in favor of (1) begins; that's when some people start warming up to the representation-with-no-X view.

Hence, the above argument says that afterimage realism, the denial of (1), inexorably *slides* into dualism, which is view (3), because views (2) and (4) are empirically implausible; call it *the Slide Argument.* We will see below that there are good reasons to reject one of the premises of the Slide Argument.

Gilbert Harman endorses option (1) (1990). He thinks that the sense datum theorists, who rejected (1), made a howling mistake. Suppose someone imagines or (what is different) hallucinates a four-legged unicorn. According to Harman, the sense datum theorists—Bertrand Russell, G. E. Moore, C. D. Broad, Edmund Husserl, H. H. Price, Roderick Firth, A. J. Ayer, and others—implied that the property of being four-legged, when we hallucinate or imagine a unicorn with four legs, is had by our imagining or hallucinating the unicorn—so we reach the absurd conclusion that our mental state or process has four legs.

> "It is very important to distinguish between the properties of a represented object and the properties of a representation of that object. … [A]n imagined unicorn [this is a represented and nonexistent object according to Harman] is imagined as having legs and a horn. The imagining of the unicorn [this is the mental process or state] has no legs or horn. The imagining of the unicorn is a mental activity. … The notorious sense datum theory of perception arises through failing to keep these *elementary points* straight." (Harman 1990, 476; my emphasis).

Ned Block presents a similar charge against those theorists.

> "[I]t is no surprise that we describe the mental image as orange even though, strictly speaking, is it not. For it is easy to slip into ascribing to representations the properties of what they represent. People who work

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> routinely with graphical representations of sounds (e.g., oscilloscope readings) often speak of them as if they had the properties of the sounds they represent—for example, being loud or high pitched." (1983, 516-17).

Although I don't want to defend everything these philosophers would have said about X, there isn't any good reason to think those eminent philosophers failed to keep straight those "elementary points" Harman describes and made the foolish "slip" Block describes. I know that they occasionally articulated claims that made it look that way—but only if one ignores the surrounding text.

> "[A] person *A* is perceiving a material thing *M* which appears to him to have the quality *x*, may be expressed in the sense-datum terminology by saying that *A* is sensing a sense-datum *s*, which really has the quality *x*, and which belongs to *M*." (Ayer 1940, 58).

However, the quality *x* Ayer writes of is understood to be highly restricted: it is a "phenomenological" or "sensible" quality. Indeed, his use of 'appears' is supposed to help here. Broad makes the restriction explicit.

> "Whenever I truly judge that *x* appears to me to have the *sensible* quality *q*, what happens is that I am directly aware of a certain object *y*, which (a) really does have the quality *q*, and (b) stands in some particularly intimate relation, yet to be determined, to *x*." (my italics; Broad 1965/23, 89; cf. Price 1932, 3).

Ayer, Broad, Moore, Price, Russell, Firth, and Husserl (and, more recently, Frank Jackson [1977, 89]) wrote a great deal about color and shape properties in hallucinatory experiences because they had good reasons for thinking that *some* properties of the object one is hallucinating—the blueness of the unicorn, not the four-leggedness of the unicorn—are actually had by the hallucination-image. *They saw that color and shape were distinctive* (which is not to say that those are the *only* classes of distinctive properties). The sense datum theory arose not because of some elementary mistake or slip but because some philosophers saw that colors and shapes are markedly different from four-leggedness.

This is not to say that the case for the existence of X is founded on some principle of the form 'If you have an experience that is phenomenally F, then something relevant is F'. Again, the main case for X is 'Well, I just *see* it, plain as day!'

The realist about afterimage X need not be a sense datum theorist, with all the accompanying epistemological baggage. Neither need she say

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

anything about hallucinations (since they are not the same phenomenon as afterimages, despite some similarities) or experiences of imagining. She's not foolishly making any grand pronouncements about *all* afterimages, as the class is highly diverse. *All* she is doing is making the modest specific claim that for the afterimage experience described earlier, something satisfies (a)-(d).

It's obvious that the reason a person comes to think of a red dagger when seeing a drawing or photograph of one is that the drawing or photo usually literally contains something—a colored patch—that is dagger shaped (in the two-dimensional sense) and red. But as we have seen, precisely the same seems to be true for the afterimage. The picture contains no real dagger; daggerness is *merely* suggested by the picture. However, the picture does contain a red dagger-shaped thing (dagger-shaped in a 2d sense) and it's in virtue of the fact that the picture is really red and dagger-shaped that it *makes* me think of a red dagger. Similarly, it's in virtue of the fact that the afterimage experience involves an orange cow-shaped object that it *makes* me think of orangeness and cow-shapedness. Or so an argument against (1) says.

To say that X is orange is not necessarily to say that it's orange in the same way a pumpkin is orange, assuming for the moment that some pumpkins are orange. Perhaps 'x is colored orange' is polysemous; this is a linguistic claim. Even if it isn't polysemous, maybe there are several truthmaking ways for something to be colored orange; this is a metaphysical claim. The critic of (1) is merely saying that something existent is orange in some way that involves instantiation of *a* color property (so it's not mere representation akin to that of how 'orange' represents orange); she is also saying that the thing that is orange is also cow-shaped (in roughly the sense that a crude drawing of a cow is cow-shaped). This point about the potential dual nature of color will resurface a couple of times below.

Critics of the representation-with-no-X view (when it is used to defend option (1)) need not hold that the view is false across the board, for all images. Contrast these two cases: (a) close your eyes and *imagine*—picture in your mind—an orange cow-shaped afterimage, and (b) actually *generate* an experience of an orange cow-shaped afterimage as described above. I do not know what is involved in case (a), the one with mere imagining. I suppose one would be generating, through the powers of one's imagination, an image M1 of an orange afterimage M2—where the orange afterimage M2 does not exist at all and the image M1 of it either is an existent brain token that isn't orange at all (or is so but only in a representationalist way akin to how 'orange' represents orange) or is

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

entirely nonexistent. Either way, all there is in case (a) is the non-cow-shaped and non-orange neurological imagining experience of the nonexistent afterimage M2. So *perhaps* in case (a) nothing in my mind or brain is orange or cow-shaped.

However, it is plain that procedures (a) and (b) are quite different. We need to do (b) in order to get anywhere on this topic. Just because the representation-without-image theory *may* be plausible for imagined images—I'm not saying it is; in the previous paragraph I just made room for that possibility—does not mean that it is plausible for our afterimage.

But forget color entirely for a moment: the case against the representation-with-no-X view is *stronger* when we set aside color entirely and focus on the shapes and spatial relations of afterimages.

Suppose you had looked at an array of three spherical light bulbs to generate three afterimages arranged as the vertices of an equilateral triangle, with each image roughly circular. The critic of (1) observes that it's very hard for most people to *look* at the arrangement of images (after they have closed their eyes and started having the afterimage experience) and deny that some things are circular and arranged equilaterally. It's much easier to make that denial when one has never looked at it. Suppose the diameter of the circles is about one third of the length of a side of the triangle. Even if the total afterimage experience *represents* all sorts of spatial properties and relations, it seems perfectly clear that there are things involved in the experience that really and truly *have* the spatial properties and relations themselves—and it's primarily in virtue of their existence that the experience does all that representing. After all, we can *see* the shapes very well.

Again, the challenge to the advocate of (1) is to account for the apparent spatial relations without saying anything that requires a really existing image with spatial relations. As before, there are plausible things to say about representation and imagination and belief dispositions, but that's the easy part. The hard part is to say things that don't, in their truthmakers, require existing images. It's so easy to slip into vague talk like 'Well, afterimage experiences are just apparent presentations of ordinary external physical objects (and sometimes extraordinary objects)'. The hard part is defending the key thesis that the "apparent presentations" don't include X. The realist of X can accept the apparent-presentation thesis—more precisely, she can accept some precisifications of it—while maintaining that the ontology of the truthmakers involved will include X. The same point holds for other ideas, such as classifying them as "illusions": the challenge is to argue that the illusion in question involves no existing image. *Some* visual illusions do not include any existing

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

images. But what makes the experience of X so philosophically interesting—and is the reason why Husserl, Broad, Russell, Ayer, Firth, Moore, Price, and others were so adamant that X exists—is that however one classifies the afterimage experience, with 'illusion', 'appearance', 'presentation', 'apparent presentation', 'representation', and so forth, I can think of one way that the claim that our afterimage is cow-shaped might be false, but as we can see it won't make a difference to afterimage realism:

Imagine you have two black circular flat disks standing on their edges on a table in front of you, so they are vertical. One of them is bigger than the other. Now imagine someone cutting each of them in half and for each disk discarding one of the halves. So now we are left with two half-circles, one larger than the other. They look like this when they are standing up vertical on the table:

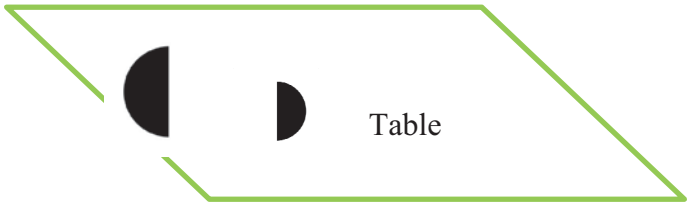

Fig. 5.1.

Now imagine arranging them so that one is a foot directly in front of you, the other is two feet directly in front of you, but they are lined up in such a way that to your eye they form an entirely homogeneous black perfect circle: the smaller half-disk is on the right, the larger one is on the left, but the larger one is far enough away from you that it looks the same size as the smaller one and lined up in such a way that they seem to form a whole circle. In this case, it only looks as though there is a *single* circular thing in front of you.

Hence, it's *possible* that the orange cow-shaped image isn't really cow-shaped but is really several images lined up to just look that way. Even so, I doubt that *our* afterimage is like that (never mind the question of whether any *other* afterimage might be like that; the class of afterimages is highly diverse and God only knows how it can be extended in highly creative ways). I'm not saying that our afterimage can't be several images lined up because it is a single two-dimensional object, unlike images lined up. In fact, I suspect the afterimage isn't two-dimensional at all, even approximately, for reasons I'll get to when examining option (4).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, even if I'm wrong, and our afterimage is really several images together, that would only mean that there are several images of various shapes. Option (1) would still be false. Henceforth, I will simply assume that if our afterimage exists at all, it is singular and cow-shaped.

Although the critics of (1) are taking the commonsensical position on the afterimage—there really is something there that's orange and cow-shaped—they need not be at all motivated by a dedication to common sense. They are not arguing 'According to common sense, there are afterimages; when a proposition is commonsensical, there is enormous warrant for it; there is little reason to reject common sense in this particular case; thus, we should reject (1)'. Indeed, they might be the type of philosopher who rejects common sense in many areas (I can serve as proof: I reject (1), I reject Moorean responses to anti-commonsensical philosophical arguments, and I even endorse some anti-commonsensical philosophical theories). To see this, consider some other views that contravene common sense.

The compositional nihilist says that there are no composite objects; trees, if they existed, would have to be composite; thus, there are no trees. The eliminative materialist says that no one believes anything. In each case there is a 'but of course' thesis. The nihilist says there are no trees but of course where you think there's a tree there is *something*: a whole bunch of merelogical simples in a tree configuration. The eliminative materialist says you didn't take out the trash because you believed the trash gets picked up today, but of course where people think there are beliefs there is *something*: a whole slew of cognitive states and processes that cause your behavior but do not have what it takes to be beliefs. These anti-commonsensical philosophers always find a substitute for the thing they are denying existence to: the nihilist substitutes pluralities of particles for trees, the eliminative materialist substitutes theoretical cognitive states for beliefs. The problem with the advocate of (1), which does not apply to the other anti-commonsensical philosophers, is that she has no plausible substitute, no reasonable 'but of course' thesis. She can try to say that although X fails to exist there is *the experience of X*, which really does exist and can serve as the substitute. But as soon as we inquire into what that experience is, as we saw above it seems that it has to involve something that satisfies (a)-(d). The lesson is that even if you are welcoming to anti-commonsensical theories, you can still find good reason to balk at (1).

For what it's worth, my experience suggests that in almost all cases it is very difficult to get non-philosophers on board with (1). When a person actually *generates* the afterimage experience, and *attends* to it, they almost

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

inevitably think it's obvious that there was something cow-shaped and orange. They are of course hesitant to admit that it's physical or that it's "objective". But the idea that there is *nothing whatsoever* there that's colored—nothing subjective, nothing illusory, nothing non-physical, nothing *at all*—is something they almost always reject. They think that although the "something" in question may be ephemeral, temporary, fuzzy, perhaps non-physical and ultimately mysterious or irrelevant, it seems as certain as anything ever gets that it was *there* when they closed their eyes, that it *existed*, and was obviously orange in some way not at all like how 'orange' represents the color orange (that's the business about orangeness being "spread out").[3] When one has the afterimage experience upon closing one's eyes, *something* is "lit up", they claim, even if it's difficult to say what it is. Upon closing one's eyes *the darkness is interrupted by a colored expanse*; simple observation is sufficient to establish that small but crucial point. Again, maybe the image isn't colored in the very same sense that an ordinary external object is colored, but it's obvious that it is colored in a very robust way—if anything, it's colored in a way more robust than that of an ordinary material object such as a pumpkin. Or so it appears.

Finally, recall that X was defined to be an object that satisfies four conditions:

(a) You were visually experiencing it from 4:00:01 to at least 4:00:10 even though your eyes were closed.
(b) It was cow-shaped (at least approximately, like a drawing of a cow) during that whole time.
(c) It was orange during that whole time.
(d) It was never red during any of that time.

In arguing against option (1) we didn't need to bring up (d), the bit about X not being red at any time from 4:00:01 to 4:00:10. All that really mattered is that X was a cow-shaped patch of color. When examining veridical color perception, claims like (d) are relevant (e.g., some theories say that objects can be two or more completely different colors all over simultaneously, so negative claims similar to (d) can be plausibly denied).

[3] The philosophers of perception mentioned earlier were just as vehement about the existence of such images (e.g., Moore [1965/1957], 134; Price [1932]: 3, 63; Broad [1965/1923]: 89-94).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

We have seen that option (1) faces an uphill battle. I don't reject it myself, but I do think there are two strikes against it. First, it is contrary to visual experience, as it seems as though that when we do the afterimage experiment we can see perfectly well that something is colored and shaped when we close our eyes. Second, the option (1) advocate has to defend not just a positive claim (e.g., 'What makes the afterimage experience so similar to the experience of seeing a blurry colored object in the dark, when it comes to color and/or shape, is completely exhausted by facts about representation') but a crucial negative claim (e.g., 'What makes the two experiences so similar does not involve any existent orange, pumpkin-looking thing', 'The apparent presentation/illusion includes no existing colored object'), and the negative claim is very hard to defend, as most attempts use terms (e.g., 'apparent presentation', 'illusory experience') that give us no reason to think there is no object X involved as a part.

Option 2: the Afterimage Is an External Physical Object

If one rejects option (1), then one is admitting that something, X, existed from 4:00:01 to 4:00:10. But that means X can't be the light bulb (or the surface of the light bulb), since it was destroyed at 4:00:01. This criticism also shows that X can't be some light waves either: your eyes are closed and there are no light waves coming from the light bulb anymore.

What about saying that one is experiencing the light bulb even though it no longer exists? So 'x is currently experiencing y' can be true at a certain time even when y no longer exists at that time.

That's fine: in some sense when you're having the afterimage experience you are "experiencing the light bulb" even though the light bulb no longer exists. But that issue doesn't matter to the ontological point under investigation. The question we're focusing on isn't "What were you experiencing while having the afterimage experience?" The question is "What is X?"

Option 3: the Afterimage Is Non-physical

Some people find themselves tempted by the idea that X had to have been a *non-physical object* that (a) was something you experienced, (b) was cow-shaped, (c) was orange, and (d) was not red. And that would prove that aspects of our sensory life are non-physical even though colored and having certain shapes!

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

For what it's worth (maybe not much), most people who have investigated these issues think this option is unlikely to be true, for several reasons:

- If the afterimage is caused to exist by physical processes, then it sure seems that it's got to be physical as well and located more or less where its physical causes are. And yet, option (3) is saying the image isn't physical and isn't located in physical space.
- If it has shape and color, then it's got to be physical. And yet, option (3) says it's not physical.
- The whole notion of non-physical aspects of the mind is fraught with difficulties that philosophers have discovered over the centuries (that I won't go over here).

Even if one isn't a physicalist, because one believes in gods or ghosts or abstract objects or even immaterial human souls, it sure seems that *when it comes to human visual sensation*, there aren't any non-physical tokens even if there are non-physical properties (so a version of property dualism is true). I don't endorse these arguments against (3), but many will.

A very different way to fill out option (3) is to hold that X is *abstract* (neither temporal nor spatial). But this is highly implausible for three reasons. First, it's pretty clear that (if X exists) it has a temporal existence: it comes into being and fades away. Abstracta are usually thought not to have temporal properties like that. Second, X is visually experienced, as per condition (a) in the characterization of X, and it is difficult to see how one could visually experience an abstract object (e.g., we may have cognitive access to numbers but Platonists don't think we *visually* experience such objects). Third, X has spatial properties and relations. For instance, it is cow-shaped in the sense described earlier; and if one has several afterimages at once, then there can be spatial relations among them. The images might not be in physical space, as the first way of filling out option (3) says, but they surely have spatial properties if they exist. Hence, it is quite doubtful that X is an abstract object.

Option 4: the Afterimage Is a Physical Part of Your Body

There are just two physicalist options for afterimage realism: the image is either an *external* physical thing or an *internal* physical thing. We already saw, on straightforward empirical grounds, that the first idea is unlikely (that was option (2)). But perhaps some physical part of your

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

eyelid or eye or CNS (central nervous system) generally was cow-shaped (again, in the manner of a drawing of a cow), orange-but-not-red, and was the thing you were experiencing: the afterimage X is just an array of cells. If so, then the second physicalist idea could work. That's option (4).

Almost everyone will object that empirical investigation shows that no part of your eyelid/eye/CNS is cow-shaped. They will also insist that empirical investigation shows that no part was orange, then red, then pink, and then violet—which again is inconsistent with identifying the image with some internal physical thing. And that seems to be the end of the matter: because option (2) is no good, we are left with (1), (3), and (4); but as we just saw (4) is no good either; so we are left with (1) and (3); hence, if we reject (1) then we are left with dualism, (3)—precisely as the Slide Argument said.

Not so fast. When you have the afterimage experience, you are experiencing—some philosophers find 'looking' and 'seeing' a little odd here—a part of your body, which is object X, and experiencing orange. It's true that if an external observer were to look at X, she would not see orange. (Here we pretend that she can look inside you; alternatively, change the example to one in which you have an afterimage experience with your eyes open, so people can examine at least your eyes.) However, this is not surprising: you are viewing X in media utterly different from that of anyone else: you are seeing (or, if you prefer, experiencing) part of the inside of your body without looking in the usual way. Due to your unique access to your own eyelid/eye/CNS, it's no surprise that no one other than you experiences the color you experience when they look at your eyes. We already know that an object can appear different colors depending on the media through which one sees it (e.g., a fish looks to be one color in the ocean and another color when brought out of the water). In fact, it's probably a stretch to say you are experiencing your afterimage "through a medium", at least in any ordinary way.

More carefully, we should not take any position on X's "real" color, provided we are being forced to use 'real color' so that an object can have just one "real" color (all over, at a specific time). To very briefly see the difficulties in thinking that all objects have "real" colors, suppose a fish looks purple to other fish in the water in which it lives but blue to us when taken out into the sunshine; suppose further that one can see the details of the fish's scales best when it is in the sunshine. If that's the way things are, then there are reasons to think it's "really" purple (as that's the way it looks to other fish in its natural environment) but there are also reasons for thinking that it's "really" blue (as that's the way it looks in an environment that allows maximal discrimination of its surface). Maybe the right

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

conclusion is that it has no "real" color. Let's not take a stand on that issue, regardless of the details of the specific example. The advocate of option (4) could do the same thing with afterimage X: it is orange to its owner and red, say, to an external person, but we go agnostic on the issue of its "real" color. The important point here is that X can be orange, at least temporarily, even though many visually unimpaired people do not see orange when they look at it at the relevant time.

So much for the objection that goes 'But nothing in the relevant body parts is orange'. It is not that much harder to figure out why other people don't see the shape you see when looking at your afterimage X. If X is part of your body, and it is cow-shaped (in the sense that a drawing of a cow is cow-shaped), then when people look at the X part of your body they should be able to detect the cow-shapedness. Can they?

Well, even if they can't, that doesn't mean they are not looking right at X anyway. Consider this array of letter 'O's:

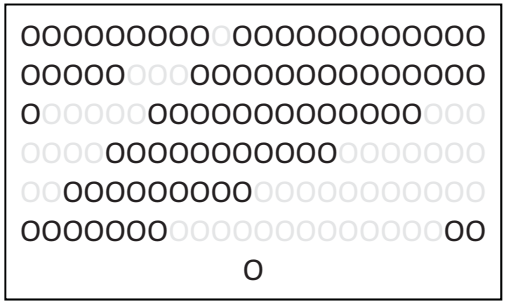

Fig. 5-2.

Let T be the triangle of yellow 'O's. If you couldn't see the contrast between the yellow and black 'O's, if you were blind to it, then although when looking at the array you would be "visually experiencing" T in one sense, as you're looking right at it, in another sense you would not because you would not isolate it from its surroundings.

Perhaps that's what happens to people when they look at your eyelid/eye/CNS. X is right there, and they are looking right at it. X is just an array of cells—just like T is an array of 'O's—and they are seeing the cells. But they can't distinguish X from its surroundings—they can't see its shape—because the main thing that distinguishes it is the fact that it, but not its surroundings, is orange, and they can't see the orangeness that's there because of difference in viewing circumstances noted earlier.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Actually, external observers probably *can* distinguish X from its surroundings by seeing its cow-shape: presumably, there's some straightforwardly physical property P, that a future scientist could find, which is instantiated in each of the cells in X but which is not instantiated in the cells surrounding X. This would be the case if the property of being orange that X has was determined by some (complex of) lower-level properties. A scientist could discover that just the cells in X have P; this would allow her to zero in on X via its shape; and perhaps she could learn of the connection between P and orangeness in order to conclude that X is orange to you. Even so, we can suppose that she is unable to "just see" that the X cells are orange, just by looking at them. Only *you* can "just see" that part of *your* eye is orange (assuming X is part of the eyes instead of some other part of your CNS). The scientist would have to learn it through testimony or scientific investigation.

Earlier I suggested the possibility that our afterimage isn't cow-shaped: what's really there are several images arranged in such a way as to give the illusion of a singular cow-shaped object. Now that I'm suggesting the afterimage is an array of cells, it should be clear that the image is not two-dimensional, even approximately (cells aren't two-dimensional). I will continue with the assumption that the afterimage is genuinely cow-shaped (like how T is triangular) and not the clever result of multiple cell arrays arranged appropriately as described earlier.[4]

It won't do to object to option (4) by saying that it is impossible to experience parts of one's eye. Larry Hardin pointed out long ago that some of the anomalous objects we visually encounter are literally in the eye.

> "It is also possible to see directly many objects and processes inside one's own eyes. They include the «floaters» in the vitreous humor, the macular pigment, the blood vessels in the retina, and «Purkinje arcs», which are probably the result of electrical discharges in the optic bundle coursing across the surface of the retina." (1988, 95)

[4] One could hypothesize that X is an internal *process*, and not an array of cells, but I have a hard time understanding how a process can be cow-shaped. Against this objection, it could be said that a cow running in a field is a process, one that has a part—in some sense of 'part'—that is cow-shaped (the cow). The advocate of 'X is a physical part of the body' option (4) could accept this idea—by modifying 'of the body' into 'of the body and its processes'—but then she has to find the part of the body that is cow-shaped, just like in the running cow story. So the requirement to find the orange cow-shaped thing has not gone away, and an array of cells (or a temporal part thereof) seems like a natural idea to pursue.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Although in that passage Hardin uses 'see', the advocate of (4) need not say that we literally *see* the part of the eye that is the afterimage. Even if 'visually see' is polysemous, it may well fail to indicate the relation a person has to her afterimage when she is, well, attending to it. This all depends on the semantics of 'see', which need not detain us. The advocate of (4) is saying that the person having the afterimage experience is "experiencing" a part of her body; the open-endedness of 'is experiencing' suffices here even if 'is visually seeing' does not.

I have been noncommittal regarding where X is: the eye, the eyelid, the brain, or what? I think that will depend on the afterimage in question; a similar point would hold for hallucinatory images (phosphenes, rainbows, holograms, etc.) that we want to be realist about. Let scientists figure out the locations. This makes option (4) hostage to the empirical facts, but this area of philosophy is hardly a priori.

An odd consequence of this way of developing option (4) is that some afterimages are *fully objective entities*. They are arrays of *cells* in one's body: physical and publically available to investigation. They exist even when not colored as in the afterimage experience. When the afterimage experience has faded away completely, so with your eyes still tightly closed you are experiencing as much darkness and as little color or light as possible, X is still there. You just no longer see its boundaries. Alternatively, one could say that X is a restricted temporal part of the array of cells, so that X exists only while you're having the afterimage; in that way we preserve the idea that afterimages are fleeting. In either case, the main thing that is subjective about the afterimage is this: only you were able to "just see" that it was orange; anyone else would have to figure it out, as described above.

Another oddity of this view is that it suggests we are highly fallible about afterimages. Indeed, we are more fallible about them than we are about familiar objects such as pumpkins. Most of us don't think of afterimages as existing independently of our experiences of them, or as material or as publically available to perception. Students typically think of them as not "really" colored, shaped, or existent. If my teaching experience is at all representative, what they "mean" is that afterimages aren't ordinary material objects that lots of people could investigate. And that's what my version of option (4) is denying.

None of this defense of (4) is intended to hold for *all* afterimage experiences, any more than observations about rabbits should be taken to apply to all mammals. The use of 'some' in this essay's title is not superfluous. Afterimage experiences (even restricted to the visual) form a highly diverse group, and there is no reason to think that one should be a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

realist about all of them (denying (1)) or adopt (4) for all of them). Neither should we necessarily apply what I've said about some afterimages to similar visual cases, such as hallucinatory images, rainbows, the sky, holograms, phosphenes, etc.

The experience of the image probably is some complicated thing involving "input" from the cortex in the attending to the image. There is still the image itself and the experience of it. My way of developing option (4) offers no insight into what the "experience" of the image is.

So far, so good, perhaps. But suppose the scientist experiences red when she looks at X, while you experience orange. That raises the question: is the red patch she experiences, P_R, identical to the orange patch you experience, $P_{O,}$ (assuming her patch exists)?

This is an issue regarding not just (some) afterimage experiences but veridical perception, which I treat in a different essay. But for now, here's an answer: P_R is distinct from P_O, since P_R is part of *her* body while P_O is part of *your* body. Hence, when the scientist looks at X, the red patch she is experiencing is part of her body, and thus not X. So we're saying that in *some ordinary veridical visual perception* the color patches we experience are in our own bodies, and not on (or identical with) the surfaces of ordinary external objects. (The semantics of 'experience' is generous and annoying enough so that 'The scientist experiences X' and 'The scientist experiences part of her own body' both come out true.) This idea is undoubtedly counterintuitive to those unfamiliar with the oddities of veridical visual perception, but those so informed should not find it counterintuitive (which of course is not to say that the idea is true).

Moreover, option (4) should not be saddled with implausible theses along the lines of 'When in ordinary visual perception one sees an orange pumpkin, what is actually happening is that (a) one is seeing or perceiving an internal physical object, and (b) one is inferring something about the pumpkin'. No, we need not accept either (a) or (b): there is no solid reason to think we ordinarily *see* the internal object in anything like the way we see pumpkins, and there is certainly no reason to think any process of inference occurs or has to occur in order to get justified beliefs in pumpkins. And of course there is no reason to wildly generalize to other aftereffect phenomena. Let's avoid bad epistemology.

I do not endorse this or any other way of developing option (4). Even so, I think it is a promising response to the afterimage conundrum. It deserves to be elaborated upon, especially since it has some significant virtues:

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

- It is physicalist at the level of tokens (unlike option (3)): afterimages and other color patches are physical things in the body.
- It is consistent with ordinary empirical observation (unlike option (1)) and science (unlike option (2)).
- It is consistent with the idea that each shade of color is a single property, so the apparent unity of shades of color can be preserved.
- Each color property may be a first-order physical property instantiated only in the eyelid/eye/CNS. So it *can* be physicalist at the level of properties too, although one *need* not accept this view, as one might hold that they are emergent, primitive, etc.
- Ordinary external objects satisfy 'x is orange' just by courtesy (i.e., in a derivative manner): pumpkins, sources of light, holograms, transparent material volumes, parts of the sky at certain times, etc. The truth conditions for 'X is orange' are exceedingly complicated, just as we have always known, because whereas parts of the eyelid/eye/CNS satisfy it in virtue of instantiating a color property (the property being one of the shades of orange), other objects satisfy it in virtue of being appropriately causally related to the instantiation of those color properties—but the causal relations are diverse, complicated, and often indirect. This is *why* we have been unable to locate color properties outside the head: there is no unitary phenomenon of orange out there.

It's a good thing that option (4) has these virtues: there are only four possible options for X, option (2) is easily refuted, option (1) is contrary to simple observation and lacks a decent defense of its crucial negative claim, and option (3) is metaphysically implausible (or so most philosophers think).

This essay is meant to merely *present* the ontological problems with color patches that philosophers of color have tended to neglect recently; I am not out to describe all the solutions or defend my favorite.

The Afterimage Paradox

We have seen that all four options for answering 'What is X?' have problems: for each one, there are serious reasons to think it just can't be right. But one of the views has just *got* to be right, no? Hence, we are faced with a paradox: one of the four options has got to be true, but for each one there are excellent reasons to think it's false.

There are seven facts that make the paradox about afterimages particularly worrisome.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

First, it can be fully presented without relying on any of the currently popular yet questionably coherent terms: ‘phenomenal character’, ‘qualia’, and ‘what-it’s-like’. Many philosophers are skeptical that much of anything truth-evaluable can be said in such terms—but as you would expect, almost none of them publish on these topics, so their skepticism goes largely unheard. By avoiding those terms and their synonyms in our presentation of the afterimage paradox, we avoid those worries.

Second, we did not need to appeal to troublesome yet ordinary terms such as ‘perception’, ‘illusion’, ‘appearance’, ‘presentation’, ‘consciousness’, or ‘awareness’. There’s a real threat that even under expert disambiguation they remain polysemous, which raises the probability that controversial arguments employing them equivocate in subtle ways. By avoiding those concepts in our arguments, we avoid those possibilities of equivocation. In addition, we didn’t have to struggle with definitions of ‘afterimage’, ‘experience’, or other troublesome terms we did employ. We stuck with a particular ordinary afterimage experience, one sufficient to reveal the paradox. Indeed, if we had desired it, we could have run through all the arguments without ever using the term ‘afterimage’. So we avoid the potential problem of having our arguments undermined by relying on flawed definitions or principles whose expression uses those terms.

Third, we didn’t have to resort to *ideal* images. For instance, although for many of the above arguments we could have used hallucinations in place of afterimage experiences, there would be no need to fantasize about the philosophically perfect hallucinations: the ones introspectively indistinguishable, even in principle, from ordinary veridical perceptions. Even if hallucinatory images are always easily distinguishable from perceptions, using nothing but introspection and ordinary effort (this is probably false, as some cases of schizophrenia suggest), we would still need to find a place for such images in our ontology, and that would be sufficient to raise the difficult ontological conundrum. And of course we could just stick with the one afterimage experience anyway.

Fourth, we didn’t have to say anything about other kinds of aftereffects, such as auditory or gustatory ones, for which theorizing is often more difficult due to lack of familiarity (both scientific and introspective).

Fifth, the question we have been addressing, ‘does X exist, and if so, how does it fit into the world?’, is independent of the answers to two of the key questions in the metaphysics of color: ‘are ordinary material objects colored?’ and ‘how are color properties ontologically related to more familiar, straightforwardly physical properties?’ Regarding the first, we didn’t assume that pumpkins, for instance, are colored and we didn’t

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

assume that they aren't colored. Regarding the second, we made no assumptions regarding whether colors are primitive properties, first-order properties that are straightforwardly physical, reflectance types, or anything else. Continuing on that theme, we made no claims about the alleged *intrinsic* nature of afterimage colors. We made no claims about the supervenience of color properties on straightforwardly physical properties. We said next to nothing about the truth conditions for 'X is orange'—other than arguing that it is satisfied for afterimages. This means that the ontological conundrum about color patches is a problem for just about everyone.

Sixth, we did not rely on any controversial epistemological principles. For centuries philosophers have argued about afterimages using questionable epistemological premises, as in the Argument from Hallucination for instance (e.g., Macpherson and Platchias [2013]). We have not done so.

Seventh, much of our argumentation need not even appeal to color, as odd as that may seem in an essay about afterimages. As we saw above, arguments regarding the shapes and spatial relations of afterimages are enough to generate the ontological conundrum regarding afterimages.

Here's what we have proven: the truth about afterimages, and as a consequence color *and shape*, is very strange and counterintuitive. We don't know *what* the truth is, but we know that whatever it turns out to be, it will be astonishing.

References

Ayer. A. J. (1940). *The Foundations of Empirical Knowledge*. London: Macmillan & Co.

Block, N. (1983). Mental pictures and cognitive science. *Philosophical Review*, 93, 499-541.

Broad, C. D. (1965/23). The theory of sensa. In Swartz, ed., *Perceiving, Sensing, and Knowing*, 85-129. Los Angeles and Berkeley: University of California Press. Originally in Broad's *Scientific Thought*. New York: Harcourt, Brace & Company, Inc.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives*, 4, 31-52. Atascadero, CA: Ridgeview Publishing Company.

Jackson, F. (1977). *Perception*. Cambridge: Cambridge University Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Macpherson, F. and D. Platchias (eds.) (2013). *Hallucination: Philosophy and Psychology*. Cambridge, MA: MIT Press.

Moore, G. E. (1965/1957). Visual sense-data. In Swartz, ed., *Perceiving, Sensing, and Knowing*, 130-37. Los Angeles and Berkeley: University of California Press. Originally in Mace, ed., *British Philosophy in Mid-Century*, London: George Allen & Unwin Ltd., 203-11.

Phillips, I. (2013). Afterimages and Sensation. *Philosophy and Phenomenological Research*, 87, 417-453.

Price, H. H. (1932). *Perception*. London: Methuen.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

—. (1997). A representational theory of pain. In Block, Flanagan, and Güzeldere, eds., *The Nature of Consciousness*, 329-340. Cambridge, MA: MIT Press.

—. (2000). *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER SIX

VISION AND CAUSAL UNDERSTANDING: PHILOSOPHICAL AND PSYCHOLOGICAL PERSPECTIVES[1]

WILLIAM CHILD

When we see an object, it causally affects us. It reflects light towards us; the light strikes our retinas; that causes impulses to be sent down our optic nerves; and so on. Without those causal processes, we could not see. But that is a scientific thesis: something we learn *a posteriori*, long after we have the concept of vision. There is nothing distinctively philosophical about this scientific thesis. And it seems clear that, when philosophers argue for or against a causal theory of vision, they are not arguing about the truth or falsity of the scientific thesis. What, then, are they arguing about? The causal theory of vision has been formulated in various ways. But there is a common basic intuition: according to the causal theory, the idea that our perceptual experiences are causally explained by the things we see is part of our ordinary thought about vision; it is an element of our naïve, pre-theoretical view of the world, rather than a feature only of a more sophisticated, scientific view.

That basic intuition has been expressed in various ways. H. P. Grice sees the causal theory as part of an attempt 'to elucidate or characterize the ordinary notion of perceiving a material object' (Grice, 1961, pp. 121-2). He concludes that the theory must not contain 'material of which someone who is perfectly capable of using the ordinary notion might be ignorant'

[1] This chapter is extracted from a longer paper, 'Vision and Causal Understanding', which was originally published in J. Roessler, H. Lerman and N. Eilan (eds.) (2011), *Perception, Causation, and Objectivity* (pp.161-80). Oxford: Oxford University Press. I am grateful to Oxford University Press for permission to publish this shortened version in the present volume.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

(Grice, 1961, p. 143). In defending a version of the causal theory, P. F. Strawson says that 'the general idea [of] causal dependence' is 'implicit' in 'the naïve or unsophisticated concept of perception' (Strawson, 1974, pp. 83, 82); and, again, that 'the idea of the presence of the thing as accounting for, or being responsible for, our perceptual awareness of it is implicit in the pre-theoretical scheme from the very start' (Strawson, 1979, p. 51). Paul Snowdon says that, for the causal theorist, it is a conceptual truth that seeing is a causal process. That implies, he says, that the causal claim can be supported by appeal to data 'that are relatively immediately acknowledgeable by any person, whatever their education, who can count as having the concept in question' (Snowdon 1981, 176). Or again: the causal theory is concerned with the 'analysis of the *concepts* of perceiving and seeing'; so a defence of the theory cannot rest only on 'arguments relying on what are, broadly, empirical considerations' (Snowdon, 1990, pp. 121-2).

The point of these characterizations of the status of the causal theory is broadly similar: they aim to distinguish the philosophical claim that seeing is a causal process from a scientific claim. But the ways in which that distinction is drawn in the passages just quoted are not equivalent. The implication of Grice's comments is that the truth of the causal thesis is known by everyone who is capable of using the ordinary notion of vision. (An elucidation of the ordinary notion, he says, must not contain material that someone who grasps that notion might be ignorant of. So if the causal thesis figures in a correct elucidation of the ordinary notion, users of that notion must know that the causal thesis is true.) Snowdon's requirement that a defence of the causal thesis must not rely on 'empirical considerations' is less demanding: for even if the truth of the causal thesis could be established without relying on empirical evidence, it would not follow that the thesis must be known to be true by everyone who grasps the ordinary concept of vision. Strawson's idea that the causal thesis is 'implicit in' the ordinary concept of perception is weaker still. It is weaker than Grice's condition: for something might be implicit in the ordinary concept without being known by everyone who possesses that concept. And, on the face of it, the idea that the causal thesis is 'implicit in the pre-theoretical scheme' is also weaker than Snowdon's requirement. After all, much of our pre-theoretical scheme – our naïve way of thinking of the world – seems to involve knowledge that is, in some sense, empirical; it is acquired on the basis of our experience of the behaviour of things in the world around us. (Think, for example, of the principles that govern the mechanical interactions of physical bodies.)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

So the characterizations offered by Grice, Strawson, and Snowdon are not equivalent. Furthermore, there is room for debate about what it takes for those characterizations to be satisfied. What makes it correct or incorrect to include the causal thesis in an elucidation of the ordinary notion of vision? What does it take for the causal thesis to be 'implicit in the pre-theoretical scheme', or to be 'a part of the very concept of seeing'? Without an answer to those questions, we do not know exactly what the causal theory of vision is claiming. There remains a strong intuition that there is room for a distinctly philosophical debate about the role of causation in our thought about perception: a debate that is not settled by the universal acceptance of the scientific thesis with which we started. But resolving that debate requires greater clarity about the intended content of the philosophical theory.

Philosophers recently have been increasingly interested in questions about the character of philosophy. What is the nature of philosophical reasoning and of philosophical theories? What distinguishes them from scientific reasoning and scientific theories? In what sense, if any, is philosophy concerned with the analysis of concepts? Is philosophy a distinctly *a priori* discipline? The questions we have just been raising about the status and nature of the causal theory of vision are instances of such questions about the status and nature of philosophical theories in general. Many of the classic writings on the causal theory of vision date from a period when it was taken for granted that the business of philosophy was conceptual analysis, and that philosophical theories are to be assessed by purely *a priori* reasoning. Philosophers nowadays tend to reject that conception of philosophy. How (if at all) and in what form does the causal theory of vision survive that change?

I shall approach those questions from two directions. In part 1 of this paper, I consider the objection that the causal thesis cannot be part of the ordinary concept of vision, since it is perfectly possible for someone to grasp the ordinary concept without accepting that seeing something involves being causally affected by it. In part 2, I reflect on the causal theory of vision in the light of psychological work on causal understanding. What light does experimental work on the origin and nature of causal thinking cast on the question, whether our ordinary thought about vision is a form of causal thinking?

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

1. Conceptual Truth and Our Ordinary Thought about Vision

On one way of formulating the causal theory, the central claim of the theory is that it is a conceptual truth that seeing an object is, or involves, being causally affected by it. And on one reading of that claim, it follows that one cannot grasp the ordinary concept of vision without accepting the causal thesis. We saw above that Grice seems to endorse that claim. But, understood in that way, the causal theory faces an objection: that it seems perfectly possible for someone to grasp the concept of vision without accepting the causal thesis.

Timothy Williamson has recently argued that there are no conceptual truths. There is, he thinks, no truth that one has to accept in order to count as grasping the concepts it contains.[2] So, in particular, there is no truth about vision that one is required to accept in order to grasp the concept of vision. Williamson's argument focuses in the first instance on grasping the meanings of words. Understanding the English word 'see', on his view, requires being a sufficiently fluent member of the practice of using that word. But someone can be sufficiently fluent in using the word 'see' to count as understanding it, even if she holds bizarre views about vision and, as a result, denies what the rest of us take to be very basic and simple truths about vision; extreme eccentricity in some elements of her use of the word can be compensated for by her normality in other parts of its use.[3] And, on Williamson's view, what goes for understanding the word 'see' goes equally for grasping the concept *see*. If someone understands the word 'see', she understands the concept it expresses: the concept *see*. So, just as she can understand the word 'see' without accepting that seeing is a causal process, so she can grasp the concept *see* without accepting that seeing is a causal process. Of course we could decide to individuate concepts in some other way; and some ways of individuating concepts would indeed make acceptance of the causal thesis a necessary condition for gasp of the concept *see*. But, Williamson argues, we would need an intellectually respectable rationale for individuating concepts that way, and it is hard to see what that rationale would be.[4]

Williamson's argument is extremely plausible. It is easy to produce actual or imaginary examples of people who plainly possess the concept of

[2] See Williamson, 2007, Chapter 4.
[3] Williamson, 2007, p. 90.
[4] For more detail, see Williamson, 2007, Chapter 4, Part 5.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

vision, but who hold views about vision on which there is no causal relation running from an object to the subject who sees it. For example, we can imagine someone accepting a 'searchlight theory' of vision. She thinks that the eye sends out visual 'rays' that range over the objects in one's environment. When an object lies in the path of these visual rays, the person's mind encompasses the object and she sees it. On this view, vision is a causal process; but the causality runs from the perceiver to the object, rather than the other way round. Or again, philosophical occasionalists hold that the objects we see do not themselves cause the experiences we have when we see things: they are only the occasions for God to produce those experiences in us. It is overwhelmingly plausible to say that the searchlight theorist and the occasionalist have the concept of vision. After all, they know what vision is; they can identify cases of seeing as well as any one else, and distinguish seeing from not seeing. They understand the causal claim about vision – which, of course, they reject. Their own false theories are clearly false theories *about vision*. Given all that, it would be implausible to say that the searchlight theorist and the occasionalist do not grasp the concept of vision. But they reject the causal theorist's claim that seeing something involves being causally affected by it. So, it seems, grasping the concept of vision does not require accepting the causal claim.

How should the causal theorist respond? The right response, I think, is to give up the idea that one cannot grasp the concept of vision without accepting that vision is a causal process. What the causal theorist should be defending is a more modest claim: that our ordinary thought about vision is a form of causal thinking. A successful defence of that claim must do three things. It must say what it takes for someone to think of vision in causal terms, or to think of vision as a causal process. It must defend the claim that we do ordinarily think of vision as a causal process. And it must show that that way of thinking of vision is part of our naïve, intuitive view of the world, rather than being a feature only of a more sophisticated, scientific view of the world.

The causal theory, on this conception, is distinct from any scientific thesis about vision. No doubt there is no sharp distinction between our naïve, intuitive view of the world and a more sophisticated, scientific view of the world. But there is a distinction. And the causal theory is concerned with our naïve thinking about vision: the thinking involved when, for example, we consider what we and other people can or cannot see ('Which of those two people is she seeing?', 'Can he see this thing from where he is standing?'), when we explain why we cannot see something ('It's too dark', 'It's too far away', 'There's something in the way'), when we explain why it looks as if things are thus-and-so, and so on. We can

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

engage in that thinking without having any scientific knowledge about the causal processes involved in seeing –about light waves, optic nerves, the visual cortex, and so forth. The point of the causal theory, on the current conception, is that this ordinary thinking is a form of causal thinking. That is analogous to the claim that our naïve thought about the behaviour of physical objects is a form of causal thinking; or to the claim that our naïve thought about the growth of plants is a form of causal thinking. In both of those cases, too, we can distinguish our naïve thinking from more sophisticated, scientifically-informed thought. In both cases, we can engage in the naïve thinking without having any relevant scientific knowledge. And in both cases, it is a non-trivial claim that the naïve thinking is a form of causal thinking.

I have accepted that someone may have the concept of vision without holding that seeing something involves being causally affected by it. The searchlight theorist and the occasionalist are cases in point: they grasp the concept of vision; but they explicitly deny that we are causally affected by the objects we see. If we hold, with the causal theorist, that our ordinary thinking about vision is a form of causal thinking – that our ordinary thought represents objects as causally responsible for our perception of them – what are we to say about the searchlight theorist and the occasionalist? There seems to me to be two possibilities. (i) We might say that, though our ordinary, naïve way of thinking about vision is a form of causal thinking, it is possible for someone to think about vision in a different way, which does not represent our visual experiences as causally dependent on the things we see. So, in particular, the searchlight theorist and the occasionalist have ways of thinking of vision that do not so represent it. An analogous position concerning our thought about physical objects would be this: 'Our naïve thought about the behaviour of physical objects is a form of causal thinking: when one object collides with another and sets it in motion, we think of the first object as causing the movement of the second; when a ball hits a window and the window breaks, we think of the ball as causing the window to break; and so on. But there could in principle be ways of thinking about these kinds of relations that did not represent them in causal terms; for example, a way of thinking that represented events of the relevant kinds as constantly conjoined without representing them as causally related'. (ii) We might, instead, take a more ambitious view. In thinking of something as a case of vision, we might say, one thereby thinks of it in causal terms. The searchlight theorist and the occasionalist have theories of vision that explicitly deny that seeing something involves being causally affected by it. Nonetheless, their basic, ground-level thought about vision still represents it in causal terms. So

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

there is a tension in these theorists' thought: their explicit theory of vision denies that it has a feature that their ordinary thought about vision represents it as having. An analogous proposal in a different area would be this: 'When one thinks of *x* as breaking *y*, one thereby thinks of *x* as causing *y* to break. Nonetheless, someone may have a bizarre theory that denies that *x*'s breaking *y* involves *x*'s causally affecting *y*. Perhaps she is an occasionalist: *x* does not causally affect *y*; it is simply the occasion for God to produce a change in *y*. Or maybe she thinks that what happens when *x* breaks *y* is this: *y* spontaneously disintegrates and draws *x* into contact with it. Such a person has the concept of breaking: she can pick out cases of breaking, and her use of the word "break" passes muster in the community. But there is an internal tension in her thought: in thinking that *x* breaks *y*, she represents *x* as causally affecting *y*; but her explicit theory of breaking denies that *x*'s breaking *y* involves *x*'s causally affecting *y*'.

This second, more ambitious, view seems right for the case of breaking; in representing something as a case of *x*'s breaking y one really is thereby representing it as a case of *x*'s causally affecting *y*. But I am inclined to think that the first, less ambitious, view is more plausible for the case of seeing. That is to say, it is possible to represent *S* as seeing *o* without representing *S* as being causally affected by *o*. My reason for distinguishing the two cases in that way is the following. Suppose the bizarre theory about breaking turned out to be true: suppose that, in cases that we ordinarily call 'instances of *x* breaking *y*', what happens is not that *x* causes *y* to break; instead, *y* spontaneously disintegrates and draws *x* into contact with it. What we would have discovered would not be that the process of one thing's breaking another was very different from what we had thought: that it did not, after all, involve *x* causally affecting *y*. Rather, we would have discovered that the cases we ordinarily regard as ones in which *x* breaks *y* are not cases of *x*'s breaking *y* at all. But things seem different for the case of vision. Suppose the searchlight theory or the occasionalist theory of vision turned out to be true: it turns out that, in cases that we ordinarily regard as instances of a person's seeing an object, there is no causal relation running from object to perceiver. Should we conclude that these cases that we ordinarily regard as instances of seeing have turned out not to be instances of seeing at all? Or should we rather conclude that, contrary to what we ordinarily thought, seeing something turns out not to involve being causally affected by it? My own sense is that this latter view is more plausible. But if that is right, then it is not true that, in representing something as a case of someone's seeing an object, one cannot fail to be representing it as a case of the object's causally affecting the person.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Some philosophers would object that, if we concede this much, then we are no longer defending a philosophical causal theory of vision. Once we allow that someone may have the concept of vision without accepting that seeing something involves being causally affected by it, the objector will say, and once we allow that someone may represent something as a case of vision without thereby representing it as involving causation, all we are left with is the claim that our ordinary, pre-theoretical way of thinking about vision does as a matter of fact represent vision as involving the causal dependence of our experiences on the things we see. And, it may be said, there is nothing philosophical about that: it is just an empirical claim about the way we think. I do not agree that the concessions I have made leave us defending a claim with no philosophical content. For one thing, the question, what it takes for a given way of thinking to be a form of causal thinking, is not an empirical question; it is a distinctly philosophical question. And in considering whether our ordinary thought about vision is a form of causal thinking, part of what we are considering is precisely that question. For another thing, the project of charting the most general features of our conceptual scheme – the project of descriptive metaphysics – has a distinguished history as part of philosophy. It is no shame for the causal theory of vision to be part of such a project.

The causal theorist claims that our ordinary, pre-theoretical thought about vision is a form of causal thinking. What can be said in favour of that claim? Consider, for example, how we tell which of two similar objects someone is seeing. We move them about, one at a time, and see which movement makes a difference to the person's experience. That procedure, the causalist says, is exactly the same as the procedure we adopt in any other case where we are testing which of two things produce a given effect. Suppose we want to know which of two switches controls the light. We press each in turn, and see which of them makes a difference to the state of the light. In that case, we are testing for the presence of a causal relation. And the same is true in the case of vision; testing which thing *S* is seeing is testing which thing is causally affecting *S*: which thing is causally responsible for *S*'s experience. Similarly, the causalist says, thinking about vision involves thinking about the enabling and defeating conditions of vision. When we think about vision, we do not just have thoughts of the form 'I am seeing *x*', or 'She is seeing *y*'. We also think about what we and others can and cannot see: 'She can't have seen the object, because it wasn't there, or it was too far away, or there was something in the way, or the room was too dark'; 'This must be the object she was seeing because this is the one that was in her line of sight'; and so

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

on. And these enabling and defeating conditions are causal conditions: they are conditions on an object's causally affecting a person. Reasoning of this sort about vision is ubiquitous in our ordinary thought. And, the causalist says, in reasoning in these ways, we are engaged in causal reasoning – just as we are engaged in causal reasoning when we think that it cannot have been the ball that broke the window because the ball is too light, or because it did not hit the window sufficiently hard, or because something stopped it hitting the window at all.

The non-causalist rejects this argument. She agrees that vision is a causal process; that, she thinks, is an undeniable empirical truth. But she denies that our ordinary, pre-theoretical thought about vision represents it as a causal process. Similarly, she agrees that, in reasoning about the enabling and defeating conditions of vision, we are reasoning about what are in fact causal conditions. But she denies that we ordinarily represent those conditions as causal conditions. All that is built into our ordinary thought, she suggests, is a set of simple principles about the conditions under which one can see things: one cannot see something if it is not there, or if it is too far away, or if it is blocked from view, or if it is too dark, and so on. We accept those principles about vision and we reason in accordance with them. But it is no part of the pre-theoretical scheme that these principles have anything to do with causation. Similarly, when we test which of two objects someone is seeing, we are in fact testing for the presence of a causal relation. But we do not ordinarily think of what we are doing in those terms.

What should the causal theorist say in response? An ambitious causalist might respond that it is just not possible to think about the enabling and defeating conditions of vision in non-causal terms; in representing them as enabling and defeating conditions of vision, one is perforce representing them as causal conditions for someone's seeing something. But I shall not defend that view. My causal theorist thinks that, though our ordinary pre-theoretical thought about vision is a form of causal thinking, it is possible for someone to represent something as a case of *S*'s seeing *o* without thereby representing it as a case of *o*'s causally affecting *S*. And likewise for the enabling and defeating conditions of vision. For her part, the non-causalist insists that our ordinary, naïve thought about vision does not represent it in causal terms. But she will agree that it is possible to think of vision in causal terms, and to do so without adopting a distinctly scientific viewpoint. For we can know that objects are *causally* involved in our seeing them simply on the basis of our naïve experience of the world, without engaging in science – just as we

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

may know on the basis of ordinary experience that moisture, light, and soil are causally involved in the growth of plants.

At this stage, the debate between the causalist and the non-causalist may seem to degenerate into an uninteresting verbal dispute about what to count as our ordinary pre-theoretical thinking about vision. The causalist agrees that vision *can* be thought of in non-causal terms; the non-causalist agrees that it *can* be thought of in causal terms; they simply disagree about which way of thinking is the ordinary, naïve, pre-theoretical way of thinking about vision. I think that view of the debate is too pessimistic. There is, as I have already said, a substantive philosophical issue about what it takes for a kind of thinking to count as causal thinking. The lower we set the threshold for something to count as genuinely causal thinking, the easier it will be to show that our ordinary thought about vision is a form of causal thinking, and the more plausible the causalist's position will be. The higher we set the threshold, the harder it will be to show that our ordinary thinking is a form of causal thinking, and the stronger will be the non-causalist's position. But we do not have a free hand to set the threshold wherever we want: there are plausible and less plausible views about what it takes for something to be a form of causal thinking. We should look for the best view of what causal thinking involves. Having done that, we may find that it is quite clear that our ordinary thought about vision qualifies as causal thinking, or that it does not. My own view is that our ordinary thinking about vision plainly is a form of causal thinking.

What, then, does it take for a kind of thinking to qualify as causal thinking – for it not merely to represent phenomena that are causal, but to represent them as causal? I have only a preliminary and sketchy answer to offer to that question. But I offer the following suggestion.

In the first place, it is overwhelmingly plausible that the concept of cause is basic and unanalyzable. That means that we cannot give a completely non-question-begging explanation of what it takes for our thinking about some domain to be a kind of causal thinking. We might say, for example, that in order for someone to represent the relation between x and y as a causal relation, she needs to represent x as bringing y about, or influencing or affecting y. But, while those formulations may be true, and while they may be helpful in reminding us what causal thinking involves, they do not give us an *analysis* of what it takes to be thinking in causal terms: for the notions of 'bringing about', 'influencing', and 'affecting' are themselves causal notions.

Second, a concept may be a causal concept – it may represent the relations it picks out as causal relations – even if those who possess the concept do not possess any general concept *cause* that they are prepared to

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

apply in every case in which they apply one or another more specific causal concept. In practice, it seems clear that children do grasp all-purpose, domain-general causal and causal-explanatory concepts like 'make' and 'because' at an early stage. But there seems to be no reason in principle why it should not be possible for a child to grasp a range of specific causal concepts – such concepts as *crush*, *break*, *spill*, *sting*, *wash*, *switch on* and so forth – and to use those concepts in thinking about phenomena in genuinely causal terms, without having any more general causal concept that she can use to classify these specific kinds of causal action or causal process as instances of the same general kind – i.e. as instances of causation.

What, then, makes these concepts causal concepts? What makes the thinking that employs them causal thinking? Strawson writes:

> 'cause' is the name of a general categorial notion which we invoke in connection with the explanation of particular circumstances and the discovery of general mechanisms of production of general types of effect (Strawson, 1985, p. 135).

On this view, what makes a concept a causal concept is just that it has to do with explaining why something happened; why an event or state of affairs occurred, or came about, or persisted; what produced some event or state of affairs; why a particular thing behaved as it did, or why that kind of thing generally behaves as it does; and so on. That is a very plausible view. And by that standard, what makes our ordinary thinking about vision a form of causal thinking is that the 'because' in our reasoning about seeing ('She couldn't see it because it was too far away', and so on) has to do with the explanation of why something happened (or did not happen). In our ordinary thought about vision, we are concerned with the occurrence or non-occurrence of natural phenomena: someone's seeing this, or failing to see that. In the same way, when we reason about the enabling and defeating conditions of vision, we are reasoning about why something happened or persisted, or why something of a certain sort failed to happen. That is enough for this reasoning to be a form of causal reasoning.

2. Psychologists' Understanding of Causal Understanding

We have been considering whether our naïve, pre-theoretical thought about vision is a kind of causal thinking. In this connection, I want to consider work from developmental psychology on questions of exactly that form, about the nature and acquisition of causal understanding. I can

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

only scratch the surface of that work here. But even a brief and incomplete comment on some psychological literature will be helpful from a philosophical point of view – as well as raising questions about some claims that have been made in the developmental literature.

In the psychological literature, the phrase 'causal understanding' is used with at least two different senses. In some cases, psychologists who consider the question, whether *S* has a causal understanding of *x*, are considering whether *S* represents or thinks of *x* in causal terms. In those debates, the question 'Does our thought about *x* involve a causal understanding of *x*?' is equivalent to my question, 'Is our thought about *x* a form of causal thinking?'. In other cases, psychologists who ask whether *S* has a causal understanding of *x* are asking whether *S* knows, or understands, the kinds of causal processes involved in *x*. To have a causal understanding of something in this second sense one must have a causal understanding in the first sense too: one cannot have knowledge of the causal processes that produce something without thinking of that thing in causal terms. But the opposite is not true: one could on the face of it think of the relation between *x* and *y* as a causal relation without knowing anything at all about how *x* produces *y*.

Susan Carey's work on naïve biology provides an example of this second use of the phrase 'causal understanding'.[5] Her aim is to show that naïve biology is a much later-developing element in our thinking than either folk psychology or naïve mechanics. These latter, she argues, unlike naïve biology, are 'core cognitive modules'. Part of Carey's argument is that someone only qualifies as having a naïve *biology* if she has a causal understanding of biological processes. And, she maintains, a causal understanding of biological processes is lacking even in children as old as 6 or 7 years old. She writes:

> Until the child has constructed an intuitive theory of how bodily processes mediate between eating and growth, or eating and becoming fat, knowledge of mere 'input-output' relations does not constitute causal understanding . . . It is unlikely that the pre-school child knows of any biology-specific causal mechanisms relevant to bodily phenomena; these may just be facts that the child has observed about his and others' bodies. Animals and people grow, the heart beats, we become sleepy even if we want very much to stay awake, etc.' (Carey, 1995, pp. 284-5)

[5] See Carey, 1995.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The child who does not know of any biological causal mechanisms, then, does not have 'causal understanding' of the relation between eating and growth.

But Carey does not think that one needs a theory of the causal mechanisms relevant to bodily growth in order to think of the relation between eating and growth as a causal relation at all. She is happy to allow that someone can think of a relation as a causal relation even if she has no idea at all of any mediating causal mechanisms. She writes, for example:

> knowledge about the relation between eating and growth. . . . may be mere knowledge of an input-output relation, such as knowledge that turning on a light switch *causes* a light to go on. Such knowledge is probably acquired through being told about input-output relations explicitly ('If you don't eat your vegetables, you won't grow into a big strong girl . . .). . . . The pre-school child has no clue as to any bodily mechanism which mediates between eating and growing (Carey, 1995, pp. 286-7 [my emphasis]).

Or again:

> pre-school children's understanding of disease, like their understanding of . . . growth and bodily processes, is limited to knowledge of input-output relations – dirt, poisons, going outside with no coat on, and germs *cause* disease (Carey, 1995, p. 292 [my emphasis])

In these examples, Carey treats knowledge of 'input-output' relations as causal knowledge. So when she says that the pre-school child lacks 'causal understanding' of the relation between eating and growth, she does not mean that the child does not think of the relation between eating and growing as a causal relation at all. She is talking about causal understanding in the second of the two senses distinguished above: knowledge of causal mechanisms.

But what about the other sense of causal understanding? Carey allows that knowledge of an input-output relation may be causal knowledge. But what makes it causal knowledge rather than mere knowledge of an association? Could there be a stage in a child's development at which she grasps that eating is associated with growth, and extrapolates that association to new cases (if A eats, he will grow; if B does not eat, he will not grow); but at which she does not think of the association in causal terms at all? If not, why not? But if there could be such a stage, what makes it the case that a child at a later stage of development is engaging in causal thinking rather than merely thinking about regularities?

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

These questions receive some treatment in an interesting literature about infants' perception of causation, which explores the extent to which very young infants perceive interactions of various kinds as causal interactions.[6] The primary focus of this work is the *perception* of causality, rather than the more general issue of what it is to *represent* a relation as a causal relation. But work on the perception of causation must take a position on the more general question. For in order to explore the extent to which infants perceive certain relations as causal relations, we must know what it takes for a perception to have causal content; and answering that question requires some answer to the question, what it takes for representations in general to have causal content.

I want briefly to explore some issues about the bearing of this work on our earlier discussion of the causal theory of vision. I focus on the overview offered in Saxe and Carey's paper, 'The perception of causality in infancy' (Saxe and Carey, 2006).

Saxe and Carey accept, for the sake of argument, the view taken by Michotte (whose work they are discussing): that we have an innate representation of cause.[7] Their own view is that it is an empirical question whether or not the representation of cause is innate. But, they think, it is a live possibility, compatible with current evidence, that 'representations with the content *cause* [are] innate' and are 'part of a central conceptual system that integrates information' provided by different sources of information about causality (Saxe and Carey, 2006, p. 163). Even if our concept of cause is innate, we can still ask what makes that concept a concept *of causality*. Possible answers to that question would include that the innate representation is a representation *of causality* in virtue of being reliably triggered by exposure to causal relations; or in virtue of its biological function; and so on. But Saxe and Carey do not address that question. So as far as their 2006 paper goes, all we are told about what it takes for someone to represent a relation as a causal relation is this: to represent a relation as a causal relation is to represent it in a way that employs one's innate *cause* representation. The main focus of their discussion is the question, what reason we have for thinking that infants do represent events of various kinds in causal terms: Do the experimental data support the claim that infants represent the world in causal terms? Or are

[6] For two early contributions to that literature, see Leslie, 1982. and Leslie and Keeble, 1987. For a comprehensive recent survey, see Saxe and Carey, 2006.
[7] Saxe and Carey, 2006, p. 148. For Michotte's work, see Michotte, 1963.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the data consistent with the hypothesis that infants represent the world only in some more basic, non-causal way?

The psychological literature that Saxe and Carey bring together does not, then, directly address the question, what it takes for our thinking about some domain to be a form of causal thinking. But it may still deliver insights that are relevant to our question. For one thing, we can infer something about what psychologists take causal representation to involve from what they regard as strong evidence for the presence of such representation. For another thing, if we are convinced by psychologists' case for saying that infants as young as 6 or 7 months old do represent their environment in causal terms, that will imply that the threshold for a representation's counting as a causal representation is relatively low. That in turn will make it easier to show that the causal theorist is right to say that our ordinary, pre-theoretical thought about vision represents seeing in causal terms.

Saxe and Carey argue that the studies they review do indeed 'suggest that young infants (by 6-7 months of age) perceive and interpret' events of various kinds causally (Saxe and Carey, 2006, p. 162). What evidence do those studies provide?

There are simple situations that adults reliably perceive in causal terms: e.g. when adults are shown a scene in which an object, A, approaches and makes contact with another object, B, and then B immediately moves off, they reliably perceive this as A's causing B to move – as A's 'launching' B. Other similar situations are not perceived by adults as involving causality: e.g. if A approaches B but stops before it makes contact, whereupon B starts moving, we do not see A as launching B; and similarly in cases where A does come into contact with B but there is a short delay before B starts moving. Taking sets of cases like these, experimentalists then ask whether infants reliably distinguish between the kinds of events that adults perceive as launching events and the kinds of events that adults do not perceive as launching events. If infants do make such a distinction, that is taken as evidence that, like adults, they perceive the relation as causal in the first kind of case but not the second.

However, as Saxe and Carey observe, the fact that infants make such a distinction is not by itself conclusive evidence. For infants might perceive the two kinds of case differently without the difference in the contents of their perceptions being a *causal* difference. 'The challenge for researchers remains to show that infants perceive these events in terms of *caused* motion (rather than merely predicted motion)' (Saxe and Carey, 2006, p. 151). They argue, however, that the hypothesis that infants do indeed perceive such events in causal terms is strongly supported when we take

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

account of further evidence. I shall mention two of the kinds of evidence Saxe and Carey cite.

First, 'infants categorize different spatiotemporal patterns together on the basis of whether they specify a causal interaction or not' (Saxe and Carey, 2006, p. 151). That is to say, infants distinguish events that adults perceive as launching events from events that adults perceive non-causally; but they do not distinguish amongst the different kinds of events that adults perceive non-causally (those where A stops before it hits B; and those where A hits B but there is a delay before B starts moving).[8] That, it is said, shows that the difference between causal and non-causal cases is in itself a salient difference for infants. And that in turn, say Saxe and Carey, is evidence that they are representing the causal cases in terms of causality.

Second, a range of experiments show that infants have a 'systematic and pervasive sensitivity to the dispositional causal status of the entities involved in the interactions' they observe (Saxe and Carey, 2006, p. 162). That is to say, their expectations about the behaviour of objects involved in events of various kinds – including the kinds of launching events described above – are sensitive not just to the objects' spatiotemporal properties, and not just to physical properties such as size and weight, but also to the kinds of objects they are. For instance, if A and B are inanimate objects, infants are surprised by scenes in which A moves towards B, stops without hitting B, and B then starts moving. But if B is a person, infants are unsurprised by that sequence of events. The obvious explanation is that infants are sensitive to the fact that people but not inanimate objects have the capacity to move themselves.[9] The fact that infants' expectations are sensitive in quite subtle ways to the effects of combining a range of causally relevant properties, argue Saxe and Carey, provides further evidence that infants have representations with causal content. This sensitivity, they write:

> bolsters our interpretation that infants are reasoning causally – they are reasoning about the causes of motion of entities, and consider that the motion of dispositionally inert objects must be caused by contact with a moving entity, and that dispositional agents are better candidate causes of motion than are dispositionally inert objects (Saxe and Carey, 2006, p. 162).

[8] Saxe and Carey cite Oakes and Cohen, 1990.
[9] See Spelke, Phillips, and Woodward, 1995.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The studies that Saxe and Carey describe are certainly suggestive. But I want to register a note of caution; do the data Saxe and Carey cite really demonstrate that young infants represent the world in causal terms?

The studies Saxe and Carey discuss do show that infants are sensitive to more than just constant conjunction: for infants distinguish constant conjunctions that adults represent in causal terms from non-causal conjunctions. And they show that the expectations infants form are sensitive to the interactions of a range of causal factors, in a fairly complex and subtle way. But does that give us compelling reason to think that infants represent those factors as causal factors? Couldn't an infant form all the expectations that Saxe and Carey describe, and be sensitive to all the features they mention, without yet representing these interactions as causal interactions? Carey warns elsewhere against what she calls the 'fallacy of theory-laden attribution'.[10] She says, for example, that it is a fallacy to infer from the fact that pre-school children distinguish animals from other things that they have the concept *animal*. But isn't it equally fallacious to infer, from the fact that infants distinguish causal relations from non-causal ones, that they have the concept *cause*? I raise this point not as a serious argument against Saxe and Carey's view but as a challenge to be answered – and as a request for more discussion and more justification. Without a fuller account of what it takes for a representation to be a causal representation, my suspicion is that they set the standards for causal representation too low.

Suppose, however, that we accept Saxe and Carey's argument for the conclusion that infants as young as 6 or 7 months old represent the behaviour of animate and inanimate objects in causal terms. What if anything would that suggest about the issue we have been discussing: whether our ordinary thought about vision is a form of causal thinking? Saxe and Carey do not address that issue. But their position, applied to the case of vision, would undercut the non-causal view. The non-causalist holds that our ordinary thought about vision involves the mastery of enabling and defeating conditions. She accepts that these conditions are in fact causal conditions; conditions for the causal production or prevention of an effect. But, she says, one can grasp and manipulate those conditions without thinking of them as causal conditions. So our ordinary thought about vision is not essentially causal. If we adopt Saxe and Carey's approach, however, that position seems untenable. The non-causalist agrees that we reliably classify instances as cases of seeing or not seeing;

[10] Carey, 1995, pp. 279-80.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and she agrees that, in doing so, we are sensitive to the interactions of a varied and complex range of causal factors. On Saxe and Carey's approach, however, that in itself is compelling evidence that our ordinary thought involves representations with causal content – that it is a form of causal reasoning. To defend her position, therefore, the non-causalist needs to set out and justify a different and more demanding standard of what it takes for a representation to be a causal representation.

My own view, as I have said, is that Saxe and Carey do set the standard for causal representation too low. But, as I argued in section 1, even when we adopt a higher standard of what is involved in causal thinking, it remains the case that our ordinary thinking about vision is a form of causal thinking[11].

[11] Earlier versions of this paper were presented at the Warwick Workshop on Understanding Perception and Causation in April 2007, and at the Catz Work in Progress Group. I am extremely grateful to the participants in those discussions, and to an anonymous referee, for very helpful comments.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER SEVEN

TEMPLATE IDENTIFICATION IN THE COMPUTATIONAL MODELS OF SELECTIVE VISUAL ATTENTION

KEYVAN YAHYA

This paper aims to address how the functional task of visual template identification has been progressed in the light of computational modelling of selective attention. To do so, a bundle of models from the past 70 years together with their turning points, conceptual elements and retrospective objectives will be surmised in a contingent manner. After presenting a preliminary introduction highlighting some of the significant endeavours that brought on the need to learn more about visual information processing, it will be depicted how the accumulation of old theories gives way to new ones that modify both of the cognitive and neural accounts their predecessors have already provided. At the next step it will be demonstrated how the continuous process of modelling template identification, thanks to the remarkable progress made in developing powerful computers, could have achieved an exceeding amount of affordability to part with the monopoly of pure psychology and instead to be gear up for hosting multidisciplinary views. Thus what comes next accounts for the way these fields are integrated concerning a variety of models that are to figure out the neural substrates, cognitive underpinnings and information processing architectures responsible for template identification since it is assumed that learning the elements of attention whose reciprocal interaction gives rise to template identification improves the cutting edge of our knowledge about object learning. Finally, we will finish our discussion by exposing some of the new lines of research as well as a few ongoing challenges ahead.

***

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Computational models of selective visual attention have been widely brought into the centre of the cognitive studies which attempt to reveal as many concealed mechanisms of the brain that are in charge of ruling visual attention as possible. Concerning previously done cognitive studies that followed the same purpose, people have started speculating about the likelihood of grasping knowledge about the existence of reciprocal relation between two cognitive functions of visual perception and attention. In regards with this possibility, a few neuroscientists could come up with some ideas which utterly intend to show the way the different cognitive functions stem from different cognitive levels that affect and interact with each other (Merikle & Joorgense, 1997).

Within this interdisciplinary field of study, for many years we have faced up some coercive difficulties which can be summed up under a question about finding an approach to figure out how attention and identification join together. So far the former studies imply that selective attention and learning have been so tightly linked together that one is likely to think of expressing each one in terms of the other. Furthermore, this kind of proximity led many scientists to build up some computational models that not only shed a light on these phenomena but also inspired scholars who are working on image processing in order to offer algorithms as efficient as possible to recognize the objects that appear on the visual field. According to Posner (1994), attention and identification are indistinctly coupled somehow: whenever an object appears on the visual field, our complex neural system starts to bring it into the attention (consciously or unconsciously) through identification process that is performed by many, say, inter-related mechanisms such as searching, orienting and filtering. Hence, many scientists have gathered a number of psychological and computational evidences in favour of the models in which attention and identification would be merged (Ullman, 2000).

In the course of our review, we should also keep an eye on one of the most intriguing problems that is firmly tightened with this attention, namely, template identification, which is regarded as the causal element of audio/visual recognition, and to become a case subtly reciprocates many brain areas including attention, episodic memory, semantic processing and subcortical areas such as limbic system (Seger, 2010). It must be noted that hereafter by attention we mean selective attention, and by template we mean visual template, so that these repeatedly applied terms will sound simpler and more straightforward. The main problem can also be recapped in a more practical way: given that a scattered visual field consists of various objects, we would like to understand how the visual information of an attended object enters the brain to be learned and identified by virtue of a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

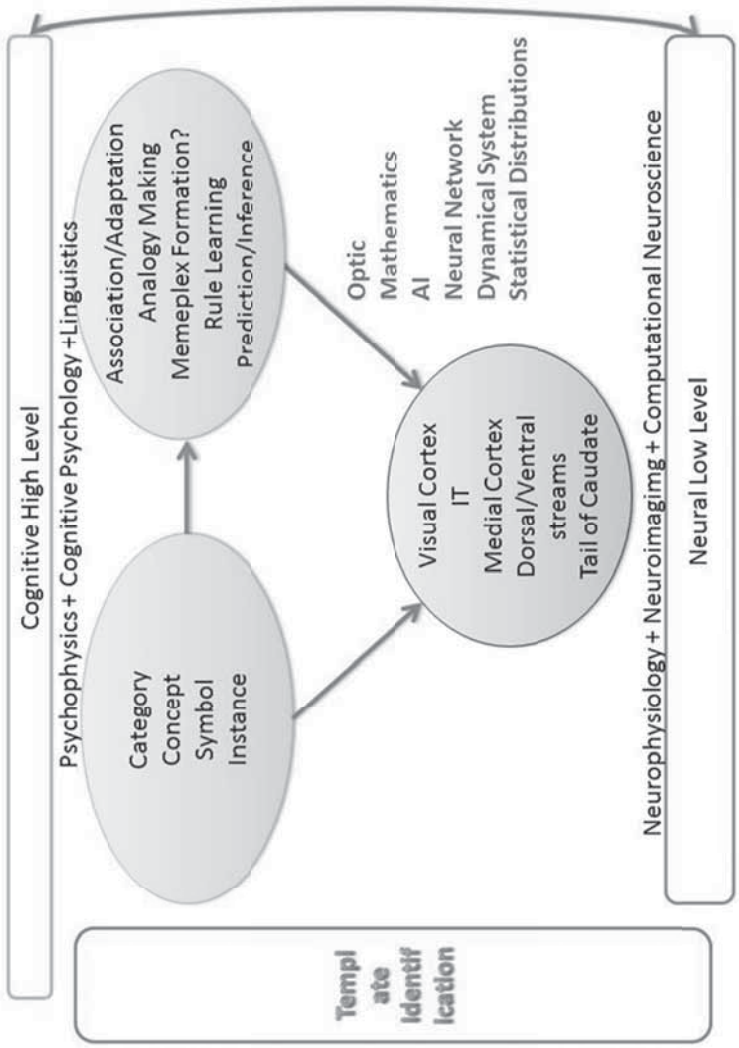

Fig. 7-1. Structural scheme of cognitive-neural collaboration in template identification, illustrating how high level cognitive processes can cooperate with their low level neural peers.

computational model of attention that entails a pair of parallel control mechanisms. Yet, few scientists have so far taken template identification

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

into account, and, to this moment, no significant study has been accomplished to set up any task that emanates from a funnelling theory and one that proceeds from a combination of information, a non-linear dynamical systems and neural network theory. Although some scholars have put forth a series of thought provoking questions in this concern, for example, about 'the gist perception' which takes place in human cognitive system or the way the brain recognizes a specific object that's been camouflaged among some other similar objects in a scattered scene (Tononi & Laureys, 2008).

Let's align the scope of the present paper, that is, summarize the problem of template recognition in the light of the various approaches, each of which should account for how template learning/identification arises from the neural activities of attention.

Functional Perspective of Visual Attention

Attention is a tangibly engrossing cognitive process emerged from a gigantic complexity constituted by neural circuits so as to help us stand out among an enormous stream of incoming information at every moment (Frintrop, 2011). Attention also plays a vital role in conducting the brain to refrain from being overloaded by hosting too much incoming information. Therefore it can be concluded that to sort things out properly the brain first needs to sample more special information regarding its current tasks, and then begins to have them processed (Huitt, 2003).

Since the middle of the nineteenth century, scientists have begun to address attention as 'spotlight', a metaphoric term that was suggested for the first time by Hermann Von Helmholtz (1850), whose ideas utterly propose that attention could occur via intentional changing of the direction of gaze, that allows to focus upon any point – whether peripheral or central – throughout the visual field. Nevertheless so long later and out of all the theories of attention inspired by this spotlight metaphor, there came a leading and well sounded theory called 'Posner Paradigm', that rests on a conception called 'biased competition', that proved quite successful in giving a description of attention (Posner, 1994). Before talking about Posner Paradigm, let's outline some essential concepts applied in building up this theory and those other ones which exploited it.

Firstly, concerning the fact that our cognitive abilities are dealt with limited attentional resources, there would be a close competition between stimuli trying to catch the resources, and secondly, winning the competition strongly depends on the attributes of stimuli and the task of attention alike (Desimone & Duncan, 1995). Since only one stimulus

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

could be winner and represented as such, just a limited capacity could be relocated to that. Seeking for a general framework to expound attention led people to take many psychophysical experiments from which some important results have arisen. In regard to the type of visual search, there would be two kinds of attentional control processes, namely *bottom-up* and *top-down,* to carry out the task of visual search, that is, finding a target among some other objects and distractors. The former (bottom-up) is a stimuli-driven and inductive attentional process, whereas the latter (top-down) is a goal-directed and deductive one (Tononi & Laureys, 2008). Cheeking on the issues of image processing would show us that a visual search task could be done in a coarse-to-fine spectrum, in the sense that lower resolution levels of an image are useful for analysing overall image texture while higher resolution levels are suitable for scrutinizing individual characteristics or specific features of an image. Analogously, we can attribute these two modalities in visual search to bottom-up and top-down mechanisms respectively.

The bottom-up process, also referred to as bottom-up expectation, takes into account 'visual saliency', that is a perceptual property of the stimulus in respect to its contrast, for example, popping a pink stimulus out of a grey visual scene including some other grey objects. Saliency is essentially related to stimulus-driven processes, and the main property of bottom-up control which does not depend on the attributes of the task and is also very fast control that could be influenced by 'figure-ground' effects (Itti & Koch, 2001). In one such process, even if stimuli are task-irrelevant they afford to catch attention. Subsequently, in a scattered visual scene, any visual search that is conducted by a bottom-up process would be biased towards the most salient object (saliency encompasses various trends like brightness, contrast, geometrical properties and so forth). Top-down expectation (prior knowledge), in contrast, highly emphasizes on visual task (instead of visual stimulus) and so it is a task oriented and biased attentional mechanism. As an example, suppose you are seeing a scattered scene in which you are intentionally seeking for a particular camouflaged object. Now, immediately afterwards the appearance of a cue pointing to your target (the object you intend to grasp), the object would be quickly attended and recognized as well. In other words, top-down process takes charge of conducting the spotlight – in the sense described above – to put it on different objects in the course of a visual search.

Both of the bottom-up and top-down processes are classified and categorized in terms of their own specific neurological substrates. According to Itti & Koch (2001) 'the expression of this top-down attention is most probably controlled from higher areas, including the frontal lobes,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

which connect back into visual cortex and early visual areas', whereas the bottom-up is triggered 'in a pre-attentive manner across the entire visual field, most probably in terms of hierarchical 'centre surround mechanism'. Eventually, it implies the same postulate we just pointed out above, holding that, at each time, only one object could be picked up from the visual field, and the other ones should remain intact. This is due to a process called 'inhibition of return' (IOR) which is another import mechanism involved in attentional deployment and prevent any selected location or processed spot from being selected again (Frintrop,2011).

Fig. 7-2. Bottom-up vs. top-down. Left: the green T seems to be the first object that quickly draws your attention. This is an example of bottom-up processing, in which your attention is captured by salient sensory information. Right: the second letters of both of the words are cut in half and so look like a same thing like two ladders of same size and shape, but top down processing allows us to read the statement and recognize the disfigured words. Adapted from Medeiros et al., 2010.

From a neurobiological point of view, the bottom-up process that is ruling the selection of the locus of attention (where to attend) is primarily controlled by the 'dorsal stream' that goes from the primary visual cortex (V1) up to the superior regions of the 'occipito-parietal cortex'. Also, it is noteworthy that object recognition occurs due to another mechanism, that is, the ventral stream, which affects top-down control. Bottom-up control is usually exerted by ventral stream that moves from V1 down to inferiotemporal cortex (IT) and from there to the visual cortex (Milner, 2012). In the recent connectionist models, both of these dorsal and ventral pathways (what and where) are usually joined up to complement each other by virtue of a parallel neural network architecture (Desimone and Duncan, 1995). Besides, Olshausen et al. (1993) have shown that those features related to the identification task are engaged with neural cluster in

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

inferotemporal cortex and ‘concerned with representating the properties of known visual shapes’. Of course, some of connectionist models of attention such as the SAIM- a computational model of selective visual attention and identification– or emergent models in general like FR-SAIM are to some extent unable of tuning to the various depositions of the retinal movements. Moreover, worth remarking would it be that "though the templates in SAIM are translation- invariant (Another important property simply means that it does not matter where stimuli (objects, templates and so forth) are going to appear on the visual field), they are sensitive to the spatial positions of parts from a particular vantage point. The SAIM is therefore "sensitive to view angle" (Heinke & Humphreys, 2003).

It would be helpful to deem the role of 'eye movements', another essential element that is of an utmost importance in modelling of attention; but a number of models have been endowed with eye movements and the rest have not. Having eye movements involved in modelling of attention not only does maintain intuition, but also helps us to examine the effect of saccadic eye movements, as well as pre-saccadic processing on template matching during visual information processing. In the technical terminology of neuroscience, the models with eye movements and the model without are referred to as ‘overt’ and ‘covert’ models, respectively (Ryu et al., 2009).

Computational Features in Building Appropriate Models

Computational studies of attention, by far and large, aim to understand the ongoing interactions of the underlying neural substrates of attention, and provide a proper account to elaborate how various tasks of information processing would be carried out whenever an object is attended on a visual scene. To do so, neuroscientists usually exploit a broad range of disciplines including mathematics, physics, computer science, neurobiology and so forth to discover as many novel facts as possible about the reciprocal relation between vision and attention. Thus, in regard to this account, we might confront different approaches that stem from different views to study visual attention even though those approaches only gain credit that ponders upon the way neural information is processed to attend a visual target. These approaches are likely to end up with the models that intend to insert an information-processing mechanism in order to observe the behaviour of visual information that goes up into short-term memory (Desimore & Duncan, 1995).

The history of cognitive science reveals that the first attempts to build the computational models of attention were made with the aid of the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

saliency conception proposed by Koch & Ullman (1985), who came up with an overt and bottom-up model to account for attention, and further deduced that attention, roughly a bottom-up speak

ing, is a cognitive based function. Their model in practice quantified and encoded the concrete notion of saliency at different locations of the visual field and then compared the amount of saliency associated with the objects of a given visual scene. We ought to notify that the vast majority of the existing computational models are endeavoured to learn as much as possible about bottom-up process, and because top-down processing stands beyond a simple topographical framework, top-down modelling of attention turned out to be a big challenge for neuroscientists. Since the top-down approach is involved with higher levels of cognition, taking such kind of low level approaches in modelling would be very less likely to provide us with any plausible knowledge (Itti & Koch, 2001).

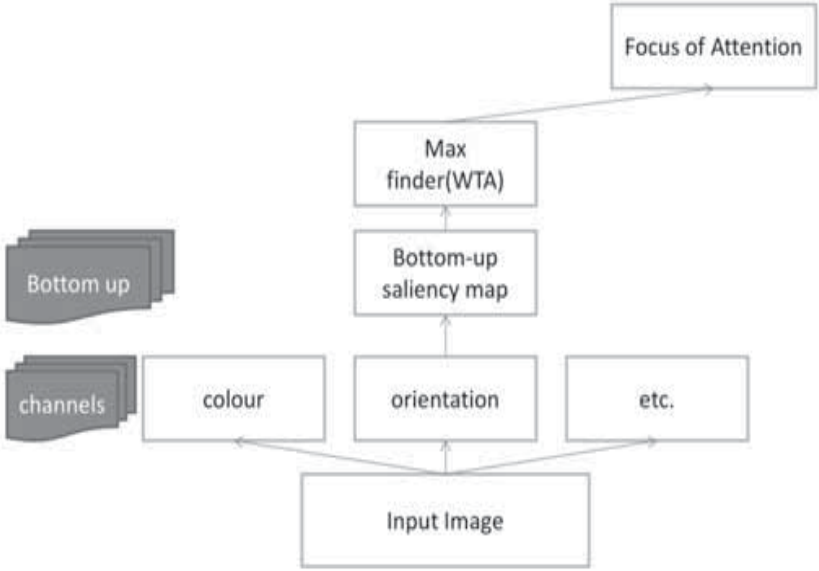

Fig. 7-3. A general architecture of a bottom-up model in which information coming into the higher level and a sigmoid function like WTA (winner take all, a neural function which detects the maximum value of what is concerned like saliency) has them summed up to terminate finally into the focus of attention.

By now, we are gradually drifting apart with the bottom-up based theories of attention and moving toward the models which benefit from the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

new features, those of cognitive higher levels. 'Posner Paradigm', given all this, is believed to be a successful framework for attention that is not only grounded upon the conceptions explained above, namely bottom-up, saliency and top-down, but adds up another novel conception of central and peripheral cues. Posner's paradigm, also known as 'cueing paradigm', suggests a three step covert model to carry out an attentional task. Attending to an object involves looking at it and putting that at the fovea (the central area of the retina with highest acuity) (Posner et al., 1978).

Therefore, according to Posner paradigm, when a peripheral target appears, subjects would move towards a central point and start responding as fast as they can. The target is cued with either a central arrow indicating the side it will appear on, or a peripheral box around the target's eventual location. Posner et al. (1978) also introduce two types of cues and use both of them in doing attentional tasks, e.g. exogenous and indigenous cues. Exogenous and endogenous cuing fit well with biased competition theory as follows: exogenous cues are triggered by bottom-up process, based on the prior expectation that salient events recur in the same part of the visual field. Endogenous cues, on the contrary, are brought into the visual field by top-down process (Feldman & Friston, 2011). One can see how the model takes a big step by engaging the problem of attention with Bayesian theory of probability. Pointing out the possibility of representing a problem that's long been believed to belong to the psychological domain in terms of logical principles is perhaps the greatest contribution of Posner's paradigm to advance neuroscience of attention.

The immediate result of his work, along with making quantitative evaluation possible, is the exposure of attention to the distinct notion of prior expectation and observation which gives way to a broader sense of inference. Attention as inference had not been suggested by anyone else before Posner's works. His theory directly influenced the neuroscientists who blossomed extensive models of attention, and, thanks to the literature, we can address some models of this sort that were constructed based on the free energy principle, which is an emergent framework embedded in Bayesian brain theory. Strictly speaking, it just suffices to say that free energy is defined as a function of data and their cause, and having that minimised would lead us to reach out the cause of our perceived data. Free Energy has been extensively used particularly in FR-SAIM (Yahya et al., 2013) and also by Feldmn & Friston (2011), both of which presumed an interpretation of free energy as Bayesian inverse or plainly prediction error.

It can be inferred that, over the course of few past decades, both motivations and methodologies to study attention have been largely varied,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and so have the goals and purposes. In retrospect we faced many turning points frequently, and a large part of these changes owed to the recent boom in technological achievements. Therefore it seems quite natural that this line of study encounters various shortages to be coped. The studies set forth and accomplished by Fred Hamker et al. (2006, 2008, 2010) are regarded as a significant progress in computational modelling of attention. His model that follows a neuro-dynamic approach takes into account some of the neglected considerable issues that examine how categorization, memory, visual search, presaccading processing and template matching can be incorporated into attention. As an example, for the first time they noted the Frontal Eye Field (FEF), placed in primate prefrontal cortex, through which saccadic eye movements can occur due to low-current electrical stimulation. The main results of their study demonstrated that FEF seems to play a significant role in temporal dynamic processing of attention through linking up to other areas such as V4, IT and TEO.

For example, when it comes to the relation between attention and memory, they demonstrated the way both semantic and episodic memories can be engaged with information processing in doing attentional tasks (Hamker & Vitay, 2008). Scientists had long seemed unable to provide any convincing explanation regarding the functionalities of attention and the role it plays in learning and categorization of templates which could be represented as a bundle of liaised concepts (Seger & Miller, 2010). It has also been shown that attention is controlled by many other underlying elements lurking down the lower neural levels such as reward pathways propagated in different areas of prefrontal cortex and tail of caudate in basal ganglia, which is now demonstrated to play many more remarkable roles than just controlling the intentional movements as it has been believed for a long time (Kolb & Whishaw, 2001).

Setting up a stunning experiment and its following results, Hamker & Vitay (2010) show that the Basal Ganglia can affect the delayed reward, and, most importantly, as the results suggest, three kinds of interrelated learning processes take place in this model as follows: the model proved successful to carry out visual information retrieval and representation tasks by attributing rewards to the objects during a learning process. The ‘stiatum’, as the input of the Basal Ganglia, “learns to represent visual information”; the ‘globus pallidus’, as its output, learns to “link striatal representations to the disinhibition of the correct thalamocortical loop”; and finally “Dopamineergic cells learn to associate striatal representations with reward, and modulate learning of connections within the Basal Ganglia” (Hamker & Vitay, 2010).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

We know that these fast and simultaneous movements of eye called saccade – originally given rise by frontal eye fields – serve us humans to reach fixation and build up a proper representation of the visual scene. Saccadic eye movements help us to select the considerable portions of the scene to rebuild it, and moreover picking the words when it comes to reading a page, through influencing fovea to use different attentional resources of the brain. But it is shown that before the onset of saccadic eye movements after a stimulus appears on the visual field, a set of successive functions should be executed to cause the brain to carry on saccadic movements. Furthermore, these functions will end up improving the localization and recognition of the saccade target. These functions put together are called 'pre-saccadic process', and comprise shifting and shrinking of the receptive fields, changing its position and compression of the visual space in a dynamic manner. Since each saccade ends up with a shift of object on the retina, the represented mapping of spatial attention should be updated around the timing of each saccade (Hamker et al., 2008).

These startling findings would however imply that the metaphor of spotlight used for attention is wrong, and also that each pre-saccade updates the next incoming saccade towards the target. Hamker (2006) first noticed all these mentioned implications to build up a multi-purpose computational model that consisted of these new assumptions to explain what and how the necessary steps that are taken during a pre-saccadic process in different areas of the brain particularly occurred in V4, frontal eye field and superior colliculus (Hamker & Zirnsak, 2006). Perhaps the most interesting aspect of this model refers to the power of prediction. It simply demonstrates that, given the fact that movement cells depict very little activities when the stimulus is presented, there is a spatially selective feedback (oculomotor) at the saccade target, which happens shortly before the onset of the next saccade and also links pre-saccade to post-saccade (Hamker et al., 2008). This model affords to predict how the mislocalization and shifting of receptive field occur concerning this oculomotor feedback.

It seems that we ought to concede a startling structure that prevailed on many of the computational studies of attention. Considering a functional role of attention, namely competition, gave way to an agreed upon rule that must be complied with in the models, that is, inserting competition between stimuli to get on the focus of attention in regard to what both of top-down and bottom-up yield. One of the most appealing efforts that proved to meet very well this requirement is imposing a mechanism that is coined as 'winner take all' (WTA) borrowed from the theory of neural networks. Roughly speaking, in a recurrent neural network WTA is a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

learning algorithm which brings about the output layer nodes to start a competition by mutually inhibiting each other while individually increasing their growing activity. If we are asked whether there is somewhere where saliency based models collide with emergent models, we could say that it must be WTA: a hierarchical mapping network containing information transmitted by input layers either as the amount of saliency or neural firing rate would be classified and processed in hidden layers, and finally arrive at an output node which is most likely to be selected by an additional output layer. Up to now, we have seen that most of the models of attention affirmatively derive out some essential elements of attention, namely competition and saliency, that any new study must be consisted of.

A considerable part of motivation for modelling attention can be traced back to an implication of a popular theory which lays down that consciousness is tightly coupled with attention, and thus a perfect understanding of attention will come about to solve the mysterious problem of consciousness. So, as far as consciousness is concerned, attention would be more likely to be thought of as some special filter so as to control the contents of consciousness. Nonetheless, many other theories of consciousness of another paradigm – for instance, consciousness as an grand illusion – cannot afford to maintain one such coupling simply because they do not ratify such things as ‘inside’ or ‘outside’ of consciousness. Moreover, there are a few models that don’t revolve around neither saliency or competition frameworks, and rather try to provide an account for attention through involving a coarse-grained Bayesian theory. Obviously, working with Bayesian theory of the brain one needs to build on any model based on prior prediction, posterior observation and dynamic improvement of prior prediction, and then let this loop go on. Putting this all in the following way likely makes sense. Given the target of attention, the model first will come up with some prior guess that emerges in other levels as the initial guess to launch the process. Then benefited Bayesian formula would lead us to grasp the target as the best hypothesis which has been emerged from ongoing interaction between prior and posterior in the light of Bayesian error elimination. Yahya et al. (2014) largely suggest a connectionist model of template identification in attention whose underpinning bears a good deal of resemblance to Bayesian theory of attention.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Conclusion

Heretofore, it seems that the models – however few in number – that privilege top-down mechanism have tuned out to be more successful than the others, whose performance is mediated by bottom-up process when it comes to explain attention. Furthermore, bottom-up based models have proved not so much efficient to account for template learning and identification. It is thought that, though having been endowed with top-down process would be a necessary condition to explain template learning and identification, it is not sufficient at all.

Scrutinizing the models working at higher levels might raise the question of whether these models could serve to understand any concealed learning process which is likely to be adhered to identification process or not. Also, worth pondering would it be to know whether these models could be reunited somehow under a certain grand unified theory which not only affords to bound them to complement one another and have them wired up together appropriately, but, as we noted, could also shed a light on the importance of the high level elements, such as competition, analogy making, prediction and so on, that play a determinate role in unfolding the whole process of template learning/identification. Furthermore, the recent achievements, thanks to ever growing brainimaging development, depict that attention is not exclusively controlled by a bundle of certain regions addressed as its neural substrates; rather there exist some other areas, in particular basal ganglia, amygdala and orbifrontal cortex whose activities are shown to have an appreciable impact on attention. Scientific theories are always in the edge of change or refutation, and this case is no exception either. Thus, taking the recent findings into consideration is very likely to result in emerging the more intricate, accurate and complete models in comparison with what we have obtained yet.

References

Feldman, H., Friston, K. J. (2012), Attention, uncertainty, and free-energy. In *Hum. Neurosci.*, 4:215.

Frintrop, S. (2011), Computational visual attention. In *Computer analysis of human behavior* (pp. 69-101).

Hamker, F. H. (2005), The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. In *Computer Vision and Image Understanding* 100 (pp. 64-106).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. (2004), A dynamic model of how feature cues guide spatial attention. In *Vision Research* 44 (pp. 501-521).

Hamker, F., Zirnsak, J. (2006), V4 receptive field dynamics as predicted by a systems-level model of visual attention using feedback from the frontal eye field. In *Neural Network* (pp. 1371-1382).

Helmholtz, H. (1850), *On the Rate of Transmission of the Nerve Impulse*. Berlin: Veit & Comp.

Heinke, D., Humphreys, G. W. (1997), SAIM: A Model of Visual Attention and Neglect. In *7th International Conference on Artificial Neural Networks*, Lausanne: Springer Verlag.

Huitt, W. (2003), The information processing approach to cognition. *educational psychology interactive*. Valdosta, GA: Valdosta State University.

Itti, L., Koch, C. (2001) Computational modelling of visual attention, *Nature Reviews Neuroscience*, 2:1-11

Kolb, B., Whishaw, I.Q. (2001) *An Introduction to Brain and Behaviour*, 2nd edition, pp. 502-510.

Medeiro, F., Verdú, B. P., Vázquez, A. R. (2010) *Top-down Design of High-performance Sigma-delta Modulators*, 1st edition, Springer.

Merikle, P. M., Joordens, S. (1997) Parallels between perception without attention and perception without awareness. *Consciousness*, 6:219-236.

Milner, A. D. (2012) *Is Visual Processing in the Dorsal Stream Accessible to Consciousness?* The Royal Society, UK.

Posner, M. L., (1994) *Attention: The mechanisms of consciousness*, Proc. Natl. Acad. Sci., USA, 91:7398-7403.

Posner, M. I., Nissen, M. J., and Ogden, W. C. (1978) Attended and unattended processing modes: the role of set for spatial location. Printed in *Modes of Perceiving and Processing Information.*

Ryu, G. G., Suh, I. H., Lee, S. (2009) *Covert Visual Attention by Object-based Selective Visual Features and Their Saliency Map*, CSREA Press, 170-173.

Tononi, G., Laureys, S., (2008) T*he Neurology of Consciousness*, 1st edition, Academic Press.

Seger, C. A., Miller, E. K. (2010) Category learning in the brain. *Annual Review of Neuroscience.* 33, 203-219.

Ullman, S. (2000) *High-level vision: object recognition and visual cognition.* MIT Press, USA.

Vitay, J., Hamker, F. (2010) A computational model of basal ganglia and its role in memory retrieval in rewarded visual memory tasks., Frontiers in Computational Neuroscience, 4(13): 1-8.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Yahya, K., Fard, P. R., Friston K. J. (2014) A Free Energy Approach to Template Matching in Visual Attention: A Connectionist Model, 12th Biannual Conference of the German Cognitive Science Society, *Journal of Cognitive Processing, 15(Supplement 1)* S75-S76.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART III:

ARTIFICIAL INTELLIGENCE: FUTURE, ETHICS AND COSTS

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER EIGHT

THE PROBLEM OF CONSCIOUSNESS ON THE MIND UPLOADING HYPOTHESIS

DIANA NEIVA AND STEVEN S. GOUVEIA

1. Mind-uploading

Since immemorial times, mankind has always sought the elixir of immortality: see the case of the ancient alchemists who believed that this product could cure all possible diseases thus prolonging life indefinitely. We can also recall the Blessed Island that fascinated so much the ancient Greeks.

But is immortality a real possibility? One of the philosophical proposals on the table is that we can do the mind uploading, upload the mind to an artificial substrate. See for example the prediction of a pioneer in the area: "the mind uploading via whole brain emulation can become a reality in the next two to four decades" (Koene, 2014, p. 98).

In 2012, a team of scientists[1] managed to make the whole connectome (the map of all the connections that neurons have with each other) of a small living being called *C. Elegans* (*caenorhabditis elegans*), and apparently to upload its 'mind' into a Lego robot.[2] The idea was to base the entire neural structure of the *C.elegan* on the computer system of the robot. Once connected, the robot went through all the paths and made all the moves that had been executed by the worm before it was analyzed. The robot started to behave and respond to the environment as the worm would do without any prior programming or human intervention. Therefore, this may be the way to get this kind of technology, but a *caveat* must be

[1] The detailed project can be found at http://www.openworm.org.
[2] To watch the video, cf. (2015)
https://www.youtube.com/watch?v=2_i1NKPzbjM.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

acknowledged: the *C.elegan* was chosen precisely for having a very low level of complexity – it only has 302 neurons. However, the human brain is at a completely different scale in terms of values: it has 86 billion neurons, each of them forms more than 10.000 connections, which have more than 100 trillion synapses each.

To achieve the same results with human brains, millions are being invested in projects[3] seeking the development of brain technology to release our minds from the biological constraints. But it is essential to understand whether this artificial substrate is really conscious. Here philosophers and cognitive scientists are divided into two positions: a) consciousness is an essentially biological phenomenon and no non-biological system can be conscious; b) consciousness is not a biological phenomenon but it has a biological structure and a causal function, therefore a non-biological system can be conscious if it is correctly organized.

For now, there is a substantial impediment: our biological body will die over time and it cannot, despite the substantial advancement of medicine in general, survive. Our brain is made up of millions of neurons that have an expiration date, in addition to all other cells in our body. But to reverse this natural event, we have to posit an artificial solution: we can, through the latest technology, make the transfer of our mind (in the broad sense of being what our brain does or creates) to a non-biological but artificial substrate (e.g., a computer). The idea is to make a technological replacement for our body, including our brain. Modeling our brain processes and transferring them from our neuronal substrate to an artificial one will be sufficient.[4]

An argument by analogy is used to defend this idea: we find no impossibility (logical impossibility, although the scientific or empirical possibility is another matter) to replace lungs and other organs and even artificial retinas. See, for example, Figs. 8-1 and 8-2: although they are constituted by different materials, both hearts do their task which is primarily to pump blood to the rest of the body:

---

[3] For example, cf. http://bluebrain.epfl.ch/

[4] A proponent of this kind of thought is Ray Kurzweil: "one just have to look at the exponential growth of computing power, efficiency and size, and then extrapolate; this is estimate based on two things: a) the estimate for the complexity of the brain and b) the estimate for the growth in computing power". He describes the mind-uploading as "scanning all of salient details [of the brain] and then reinstantiating those details into a suitably powerful computational substrate. This process would capture a person's entire personality, memory, skills and history". (Kurzweil, 2006, pp. 199).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

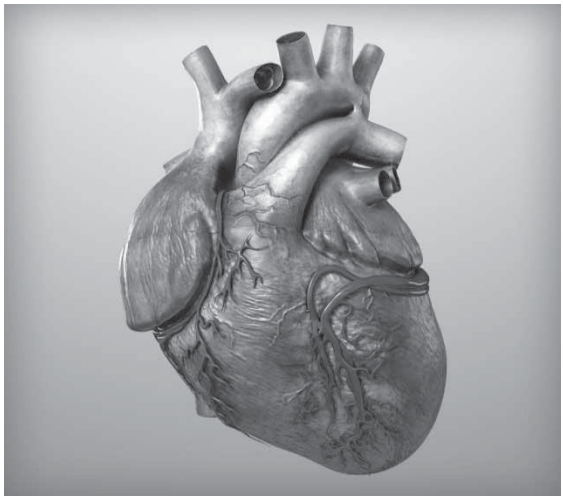

Fig. 8-1. natural heart[5]

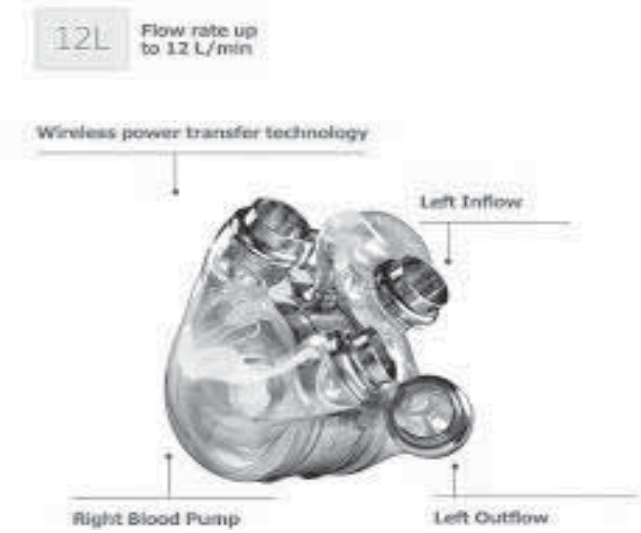

Fig. 8-2. artificial heart[6]

[5] Cf. http://cliparts.co/clipart/3586294

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

If we can replace parts such as the heart, why not replace the whole body, including the brain, with a technological equivalent?

The question placed here seems to anyone with a basic philosophical knowledge to be a clear error of reasoning: namely, the formula thus postulated commits the fallacy of composition that can be formalized by:

(1) *x* is constituted by *y* parts;
(2) *x* has *z* properties;
(3) Therefore, *y* has *z* properties.

Thus we will have to find a better argument for the possibility of mind-uploading.[7] Note that, curiously, and oddly enough, the physicalists themselves (strictly speaking, the computacionalists influenced by functionalim) are the ones to make this suggestion. After all, they argue that dualism is absurd and, as such, the mind and the brain are one and the same. But the idea of transfer will necessarily lead to a dualistic conception. See the argument:

(4) Advocates of the possibility of mind-uploading are physicalists/materialists;
(5) The physicalist/materialist does not conceive the mind as separate from the brain/body;
(6) The theory of the mind transfer takes necessarily the mind as different from the brain/body;
(7) Therefore, the proponents of mind-uploading cannot believe the theory of the mind transfer.

The basis of the argument intends to show that there is a contradiction in terms: who posits the possibility of mind-uploading derives that position from a purely physicalist/materialistic framework about the mind. It is commonly asserted that both the type identity theory (and its token identity variant) and the functionalism argue that mind and brain are one and the same substance (they are therefore monistic). But the transfer

[6] Cf. http://www.ele.uri.edu/Courses/bme281/F14/2_ScottG.pdf

[7] Anders Sandberg and Nick Bostrom examined the engineering problems that mind-uploading could bring in *Whole Brain Emulation: A Road Map* (2008), where it is argued that the detailed knowledge (how the organized structures of the brain's cortices respond to input), rather than a functional knowledge (or understanding – how the brain is organized), should be sufficient to emulate the all brain (Bostrom & Sandberg, 2008, p. 8).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

thesis seems to make use of a dualistic conception of the mind: for something to be transferable, this thing has to be different from its spatial and physical location.

This argument can also be found in John Searle, who argues that Artificial Intelligence thinkers, who so heavily criticize the Cartesianism for its dualistic conception, don't realize that they make the same mistake: "strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter" (Searle, 1980, p. 430).

Now, do we have a good argument? Although it is certainly plausible that most mind-uploading advocates are physicalists, the truth is that there may be others that are not, thus saving the position of an impossibility.

David Chalmers supports the possibility of mind-uploading, although he defends a panpsychist view of the mind (and appeals to the notion of "organizational invariance", which holds that any two systems with the same organization have the same conscious experience).

We can thus object, rightly, that functionalism needs not be physicalist: in the paper "The Singularity: A Philosophical Analysis" [8], David Chalmers defends exactly this possibility.[9]

While we will focus on this topic at the final part of the paper, an argument first presented in favor of the singularity – a term coined by John von Neumann, describing a future period of humanity when technology will change human nature (cf. Kurzweil, 2006, p. 7) – will be important for the question under debate, and that is the Emulation Argument based on the possibility that we can emulate the human brain:

(8) The human brain is a machine;
(9) We will have the ability to emulate this machine (before long);
(10) If we emulate this machine, there will be AI;
(11) Therefore (absent defeaters), there will be A.I (before long).
(Cf. Chalmers, 2010, p. 15)

This argument is also defended by Ray Kurzweil, Google Engineering Director, who made some quit accurate predictions [10] through a very

[8] Cf. http://consc.net/papers/singularity.pdf

[9] However, such a view continues to seem to assume a dualism for some authors. See, for example, Massimo Pigliucci's tough words: "(…) substrate independence of the type envisioned by Chalmers implies a form of dualism that should be unacceptable in modern philosophy of mind" (Pigliucci, 2014, p. 119).

[10] Some previsions are quite suggestive: at the end of the 2030's the mind-uploading program will be officially concluded and available with no flaws; in

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

simple method: just looking at the exponential growth of computing power, efficiency and size and then extrapolating. But how can we defend the premises (and the conclusion) now referred?

Chalmers provides us with several reasons to defend each stated premise: we know (8) through the information about biology and physics.[11] All our body organs seem to be a machine: the body is a complex system composed of self-governed parts which interact in a self-governed way. As such, the brain is no exception.

(9) follows from the statement that microphysical processes can be approximately arbitrarily simulated and that any machine can be emulated by approximately arbitrarily simulating microphysical processes.

There are various ways by which we can deny the possibility of AI: (a) in 1961, J.R. Lucas argued that for reasons related to Gödel's theorem, humans are more sophisticated than any machine or computer; (b) H. Dreyfus (1972) and R. Penrose (1994) argued that human cognitive activity could not be emulated in any computing machine; (c) J. Searle (1980) and N. Block (1981) also argued that even if we could emulate the human brain it wouldn't necessarily follow from that that this emulation had an intelligent mind. However, for Chalmers all these objections are not consequent, because all (a), (b) and (c) only maintain the possibility of AI being negative if we consider a purely classical computing framework (cf. Chalmers, 1996, chapter 9).

This way, Chalmers clears the ground in order to be able to analyze the philosophical possibility of mind-uploading.

He will expose three possible methods for that:

(M1) Destructive uploading – as the name implies, this method involves the destruction of the biological mind. For example, we could freeze the brain[12] and analyze its structure layer by layer,

---

2045 singularity will be up in its plenitude – artificial intelligence will overcome any life form; and even the projection that in 2099 machines will have the ability to create computers with the size of planets. (cf. Kurzweil, 1999 and 2005)

[11] An argument that defends that the physical conception of the Universe itself requires singularity can be found in Tipler, 2012, pp.183-193. There it is defended that a «universe collapsing contradicts the second law of thermodynamics, unless an infinite series of "Kasner crushings" happen; no carbon-based life could survive to the collapse, so it is required for artificial intelligence to happen».

[12] It is noted that these issues are not only important for our philosophical conception of the world, but it is already having consequences in practical everyday life: there are many new companies offering such a service; for a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

building a detailed map of neural connections. This information would then be used to construct a computer model of the brain functioning, destroying the original brain during the process, creating a new, totally artificial brain.

(M2) Another method would be a gradual[13] upload: the original copy would be gradually replaced by functionally equivalent elements (through, for example, nanotechnology – nanotechnological devices could be placed along the nerve cells whose function is to record its activation and use this information to simulate the neuron behavior). This would give a similar functional construction of the original neuron. Once we had all the "copied" neurons, we could destroy the original ones and the functional analogue one would take its place; then we would repeat the method to all neurons until we had a complete copy of the brain.

(M3) Finally, we have the non-destructive method to the uploading[14]: here we would retain the original copy. Through brain-scanning we would make a dynamic map of how the brain works, without destroying it, and build up one functional analog.[15] (cf. Chalmers, 2014, p. 103)

Then, two fundamental problems arise:

(P1) The problem of consciousness: would the uploaded mind be conscious? Would it experience the world as the original one?;

(P2) The problem of personal identity survival: assuming this new mind would be conscious, would the person survive the process? We will leave this interesting question aside here.

---

considerable amount of money it is possible to freeze the brain, so that in the near future the person's mind can be uploaded. See, for example, the report of an actual case of this type (cf. Hendricks, 2015).

[13] For example, Sam Bradford presents in his article "The framework for approaches to transfer of the mind's substrate" this idea: the idea here is that the human brain (in which your identity is currently "housed") could be replaced by neuroprosthetics. This is already being done to some extent, with things like cochlear implants and artificial limbs replacing the input and output channels of the nervous system (see Bamford, 2012).

[14] Bostrom proposes this type of method (cf. Bostrom, 2008, p. 40).

[15] These are just possibilities whose discussion is in the logical possibility value.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

For Chalmers the only method that seems, in principle, able to deal with the problem of consciousness is the (M2). He argues that:

(12) If the parts of the brain are gradually replaced by a functional isomorphic, our conscious experience will: either a) be lost suddenly; b) gradually disappear or c) be maintained over time;
(13) Sudden loss or fading are not plausible; conservation is;
(14) Therefore, it is likely that our conscious experience remains during the process of gradual replacement;
(15) The conservation of conscious experience is only compatible with the functionalist conception;
(16) Therefore, the functionalist conception is probably the correct one and the preservation of consciousness (and personal identity) via mind-uploading is plausible. (Chalmers, 2010, pp. 42-47)

But how can we argue that, for example, the gradual replacement of parts, even though they are functionally the same, does not cause the fading of consciousness? Chalmers defends at this point that the hypothesis (12a) is unlikely because we can always replace a copy to a more fundamental level: instead of replacing a neuron and cause the loss of consciousness, we could avoid it by replacing the molecules of neurons, and so avoid losing consciousness. We could again say that this would cause the loss of consciousness, but at this point we can say that neurophysiology will easily deny us that: after all, the loss of molecules (and even complete nerve cells) is constant in our brain, without this necessarily leading to the loss of consciousness.

Since the hypothesis (12b) is denied because it implies that, although the copy is functionally equal to the original substrate, consciousness would disappear when the upload of the different parts of the brain occurred. Such a hypothesis goes against the functionalist assumption that will be defended.

Thus, to defend premise (13), Chalmers cautions that all partial uploads will be fully conscious, since new elements are functionally identical to those replaced elements (Cf. Chalmers, 2010, p. 46). But such an assumption can only be made if we consider that a computational theory of mind is correct. Now it is known that Chalmers is against this kind of theory. But, as we have noted above, the philosopher sees no problem in assuming that a computational theory of mind is the best one if it is not connoted to classical concepts of computing, and is understood in a broader sense – taking computing as a framework yet to be discovered. It

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

follows then that Chalmers finds probable that the functional/ computational theory of mind is correct:

> My own view is that functionalist theories are closer to the truth here. It is true that we have no idea how a nonbiological system, such as a silicon computational system, could be conscious. But the fact is that we also have no idea how a biological system, such as a neural system, could be conscious. The gap is just as wide in both cases. And we do not know of any principled differences between biological and nonbiological systems that suggest that the former can be conscious and the latter cannot. In the absence of such principled differences, I think the default attitude should be that both biological and nonbiological systems can be conscious. (Chalmers, 2010, pp. 36-7)

If (13) is true, the rest of the argument works and the conclusion (16) will demonstrate that the mind-uploading is a real possibility.

2. Critics

However, this argument seems to have a fundamental problem: it assumes that the computational functionalist theory of mind is correct. Thus, demonstrating that this assumption is wrong we could destroy all the arguments so far employed. Formalizing the argument:

(17) The mind-uploading is possible only if the computational theory of mind is correct;
(18) The computational theory of mind is not correct;
(19) Therefore, the mind-uploading is not possible.[16]

For our argument[17] to work we have to give reasons to support that premise (18) is true. For that, we will (A1) present an adaptation of

[16] Inspired by Pigliucci's reflection (cf. 2014, pp. 119-130).
[17] Although we have room to work with such a subject, a philosopher who, in principle, accepts the mind-uploading thesis as possible presents empirical reasons to show that this is unlikely, due to the "combinatorial explosion - refers to the large amount of possibilities arising when one doubles the quantity of combinations" (Dennett, 1991, 5n). Although such arguments are presented for the idea of incorporating a brain-in-a-vat into a machine, the same conclusions can be drawn for this case.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Searle's Chinese Room argument applied to functionalism[18], and we will (A2) present an adaptation of the Philosophical Zombies Argument.

But before that we will define what we understand as a computational theory of mind.

3. Functionalist Computational Theory of Mind

According to the functionalists, the mind is not a thing or a different ontological substance from the matter. Rather, the mind is a kind of pattern with information constituted by the functional relationship between various elements (e.g. neurons). The functionalist belief is that if we can replicate these functional relationships, then we can replicate the mind (or brain). In addition, functionalists criticize the type identity theorists for being too restricted: only a very specific circuit can be activated when, for example, we perceive a red object. Thus, they postulate that the mind is multiply realizable. Hence, the function can be replicated by any medium, if its structure is equal to the biological structure that is replicated (neurons, glial cells, neurotransmitters, neural networks, etc.). Now, the supporter of a Computational Theory of Mind (CTM) holds that the mind is a special kind of computer (see Horst, 2009).

But is our whole brain really computable? A first influence of this theory comes from Alan Turing, a British mathematician who, when trying to solve the tenth problem[19] presented by David Hilbert in 1900 at the famous conference of the International Congress of Mathematicians in Paris, the Decision Problem, developed a precise mathematical notion or definition of computation. Informally, something is computable when: a)

[18] However, there are several types of functionalism; the one we are here addressing has the name of "Turing Machine functionalism", originally designed by Hilary Putnam. This view sees any creature with a mind like a Turing machine (so it is based on a finite state machine, a machine that can only have a finite number of states at the same time, the transition from one state to another is controlled by a look-up finite table – a finite program that specifies, for a given state in which the machine is located, which state will be the next to run a universal Turing machine that can emulate the behavior of any other Turing machine. (cf. Mandik, 2014, p. 96)

[19] Is there an algorithm which can *a priori* demonstrate if, given a mathematical sentence, it can be logically deductible of a certain axioms set? If it is deductible, it is true and we've proven the theorem; if it's false, its negation is true (cf. Teixeira, 2004, p. 57).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

there is a set of sequential instructions (an algorithm) and b) when they are followed by a machine or a person, the task is completed.

Another important notion is what would become known as the "Turing machine"[20]: a mathematical object with: a) memory – a tape on which symbols can be read, written and amended); b) a set of symbols; c) a memory bus – a scanner which reads and writes on the tape; and d) a set of instructions. If we have these elements, we can accomplish, in principle, a computation of any task.[21]

Such framework would give hope to the functionalist program in finally being able to solve the hard problem of consciousness and create a true artificial intelligence. However, it was Turing himself who, by postulating the Halting Problem[22] (see Teixeira, 2004, p. 59), pointed out the fundamental limitation of this research program.

Nevertheless, some authors advocate a pan-computationalism, but, in their doing so, this information does not help us at all and turns the claim itself futile. Consider: to say that everything is computable doesn't help us to understand the difficulty that the (conscious) mind has caused to contemporary science. Pigliucci rightly (in our view) points out that this idea is due to a dubious interpretation of the Church-Turing thesis.[23]

Another problem arises when we want to understand what kind of computer our mind is: is it digital (binary system – symbols representation) or analogical[24] (the representation has similarities to the

[20] Given the fact that this is not the main subject of this paper, we refer a detailed analysis of it in Ravenscroft, 2005, p. 89.

[21] It was the Austrian John von Neumann who implemented this model physically, creating digital computers.

[22] According to João Fernandes Teixeira, Turing reasoned as follows: "(…) if there is an algorithmic procedure to solve a particular problem, it can be represented as a Turing machine and therefore, this procedure is necessarily finite, that is, we before a Turing machine whose data processing at a certain time stops. Not to stop means being in a situation of non-algoritmicity or incomputability" (Teixeira, 2004, p. 58).

[23] In the author's words: "Turing's version of the thesis says that logical computating machines, which eventually became known as the Turing machines, do nothing which can be described as a rule of thumb or purely mechanical ("algorithmic"); the Church version says that function of positive integers is effectively calculable only if recursive. None of the above implies the sort of much stronger declarations that have been made by computationally inclined philosophers of mind" (Pigliucci, 2014, p. 123).

[24] An analog computer is contrasted to the digital computer. The 'analog' expression has several meanings: a) involves a contrast with digital systems, where

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

represented object)? Or is there still a third option: a quantum computer?[25] Does it work by a serial architecture – with only one processor or CPU – or by a parallel architecture? There may also be another option: it may be a mixed model of interaction between the analog and the digital, in which the mind is an emergent property of the analog electromagnetic field of the interaction of neurons (working digitally) produced in the brain; these emergent properties are analog and, therefore, they are not computable. The neuroscientist Miguel Nicolelis, who supports this position, gives the example of monkeys repeatedly performing the same arms movement: when studying the pattern of neuronal activation and trying to find a pattern when recording neuronal activity, the active neuronal structure is always different – this happens precisely because there is an interaction of the electromagnetic field which unsets the linearity of the digital process. (Cf. Nicolelis & Cicurel, 2015, p. 19)

Interestingly, many cognitive scientists have been arguing that the nerve cells, neurons, are small computers which compute – we have information transmitted between neurons. The electrical firing of the neuron is like a digital computer: see the works of Warren S. McCullough and Walter H. Pitts[26], who used a formal logical system to describe the neural activity. Despite the authors' consideration that their studies are only an approximation model, many of their followers took these models seriously.

However, although there are similarities in the inputs/outputs form, the truth is that neurons are always being redefined in milliseconds, while a

---

"digital" means that individual circuits are only capable of a finite number of discrete states – e.g., the numeric value of 0 or 1. Analog in a broader sense means only b) "non-digital", and applies to systems whose components are capable of continuous states – for example, numerical values that represent all real numbers from 0 to 1. However, it should be noted that "digital" is not connected to binary: a digital system – an n-valued system (e.g., 0, 1 and 2) also counts as digital. (Cf. Horst, 2015)

[25] The quantum computing program seeks first to distinguish the classical concept of information (in bits – 0 and 1) from the concept of quantum information (based on intrinsic randomness, the uncertainty principle and the entanglement that form the qubits – we define 1 and 0 according to our choice, our observation, not knowing what could be a different choice – this randomness comes from the ability of quantum states to be overlaid: only when we can measure them we have a defined state). The main advantage of this type of computing is not being required to perform calculations sequentially as in classical computing.

[26] Cf. McCulloch & Pitts, 1943.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

computer is not. Furthermore, given von Neumann's architecture, a computation has the following formulation:

Now the problem is that the natural activity of a neuron is too far from this simplicity: a single neuron may receive up to 1000 inputs at once and, depending on the sum that the cell attaches to incoming loads, you can send up to 1000 outputs to another cell.[27]

This way, on the one hand, neurons appear to be digital: either they trigger or not. But on the other hand, they also appear to be continuous: they are firing over time and they are modeled by differential equations – thus representing a continuous function. This seems to lead to a mixed model between the digital and the analog (cf. Nicolelis & Cicurel, 2015, p. 11). In addition, parallel distribution model of connectivism seems to be more similar to the way the brain functions than the classical serial model.

Moreover, if we observe several biological creatures, their exact perception circuits, for example, are all different. They are constructed in a different and dynamic way. Now, that does not happen in a traditional computer: all elements must be in the same place, connected by the Turing model – if any of the elements fails, the entire circuit fails.

Nevertheless, the cognitive science program of study is based on this assumption: for example, a cognitive scientist who wants to study the ability of man to learn a language will seek to, during the research, create a computer program (algorithm) and compare this artificial program with the response (speed, efficiency, etc.) that a normal person would give to learn that language. The idea is to make reverse engineering[28], discover what the brain code is and put it in another artificial substrate.

Another important idea is that the mind is compared to a computer (abstractly thought); or rather, the mind (or the brain) is a kind of

[27] Still, by using supercomputers, computer scientists have been developing an attempt to simulate the human brain activity. For example, a Japanese team managed to simulate the operation of a brain for a second, using more than 82,000 processors. 1.73 billion virtual neurons and 10.4 trillion synapses were recreated, each containing 24 bytes of memory (Cf. Neal, 2013).

[28] This is the main focus of the current projects of the studies of the brain, and perhaps the main theoretical reason for the limited success of the same. But doesn't this idea seem to imply that the brain would have been "engineered" in the first place? (Cf. Nicolelis & Cicurel, 2015, p. 73)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

computer[29]. Thus, it is important to note that a computer recognizes and manipulates symbols taking into account only their syntactic properties (this is what gives it efficiency and allows it to be universally programmed in any physical system). Exactly because of that, a computer does not deal with any kind of semantics, although it may, following the syntactic properties, respect the semantic properties as the true value. But is it enough to explain consciousness? The aim is to discover the brain connectome[30] and replicate it in detail in purely inorganic artificial cells. Joining the molecular matter (neurotransmitters), we could have the fully replicated brain activity. This seems to make sense only in the functional brain correlates. But could the functional part produce consciousness?

Consider then two of the most influential arguments against the theories of functionalist intuition with direct consequences for the Mind-Uploading thesis. Recall: the aim is to show that (18) is correct, so that our solid argument.

3.1. The Philosophical-zombies Argument

The first objection (O1), an adaptation of the P-Zombies Argument, needs a prior clarification of its first original version against the physicalist theory.

This argument relates to the logical possibility of a 'zombie world', a world which is exactly like ours in every aspect except in one: it lacks phenomenal properties, conscious experiences. So my twin zombie living in that zombie world would be molecule by molecule just like me but without any kind of conscious experiences. The logical possibility of p-zombies used by Chalmers would, then, demonstrate the irreducibility of mental states to body states. If p-zombies are really possible, then phenomenological states are not identical to physical states – and so, type-A materialism is wrong.

The argument is as follows:

(20) a state of consciousness *C*, such as the taste of chocolate, cannot exist without the same exact state *C*;

[29] We purposely left out the two original theoretical Computational Theories of Mind, including Jerry Fodor's (1975) and Hilary Putnam's (1960), because we believe that they subsequently argued that a CTM cannot be a complete theory of mind, precisely because many of the mental processes do not seem to be computable (see Fodor, 2000).

[30] For more details, cf. http://www.humanconnectomeproject.org/

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

(21) we can conceive of p-zombies;
(22) p-zombies show that it is possible that a physical state *F* occurs without conscience of a state *C*.
(23) According to the Identity of Indiscernibles, "for every property *F*, if the object *x* has *F*, if and only if the object *y* has *F*, then *x* is identical to *y*".
(24) Thus, if *C* can only occur with *C,* and *F* can occur without *C*, *F* is not equal to *C* (i.e., the physical states are not equal to mental states).

In short, if p-zombies are possible, then the physicalism which supposes that mental states are reducible or are equal to physical states is false.

To make this conclusion, it will be necessary to have a strong argument to prove the premise (21). We specify the argument like this:

(25) P-zombies are conceivable.
(26) What is conceivable is possible.
(27) Then p-zombies are possible.

This argument is then based in two central premises. But how does Chalmers assert both premises (25) and (26)?

For the philosopher, the conceivability of p-zombies is: *prima facie* (at first sight), *ideal* (that is, it is rational), *primary* (it is primarily or epistemologically conceivable because it can be, in fact, the case – it can be totally *a priori*), and *negative* (it is not *a priori* rejected) or *positive* (not only it is not *a priori* rejected but it also can be, in fact, *a priori* be the case). Such conditions to the conceivability turn possible a positive possibility.

These kinds of conceivabilities are the most interesting ones: these are the guide to the most interesting kind of possibility – the primary possibility. That means, it can exist a metaphysically possible world which satisfies the hypothesis when considered as real (the existence of exactly equal humans without *qualia* is possible, showing those *qualia* are something extra).

Secondly, for the premise (26) the philosopher argues that conceivability is a guide for the possibility, being an advocate of the "Conceivability Argument" (cf. Chalmers (2002), “Does Conceivability Entail Possibility”)[31]. That is, if we can conceive a zombie-world case, that

[31] Cf. http://consc.net/papers/conceivability.html

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

only possible world case suffices to conclude that this possibility means physicalism is not *necessary*.[32]

Adapting this argument to functionalism, we only have to change the implicit notion of physicalism for functionalism, obtaining something like:

(28) If functionalism is true, then it is impossible for two beings to be exactly functionally equal but different in the aspect of one of them being a p-zombie.
(29) It is conceivable for two beings to be exactly functionally equal but one of them being a p-zombie.
(30) If something is conceivable, then is possible.
(31) Therefore, functionalism is false.

This way, we show that premise (17) of our argument above, which held the computational theory of mind, influenced by the functionalist conception, is wrong, mining the possibility of mind-uploading advocated by Chalmers, and consequently the possibility of living forever, jumping from our physical support to a hardware, but keeping our mind intact.

However, it is possible that the p-zombies argument doesn't work, not undermining premise (17) and Chalmers' version of mind-uploading.

Daniel Dennett is one of the most famous philosophers to reject the idea of p-zombies, having written the paper "The Unimagined Preposterousness of Zombies". He alleges that those who are the "zombie friends" "invariably underestimate the task of conception (or imagination), and end up imagining something that violates their own definition" (Dennett, 1998, p. 172).

To address this issue Dennett coins the term "zimboe". Zimboes are zombies with higher order informative reflective states, such as beliefs. For Dennett only these zimboes would be sufficiently similar to human beings to the point they would be mistaken as normal beings; not p-zombies but just zimboes would pass the Turing test, for only they would be able to have higher order reflections necessary to make it possible for a behavior to be so similar to a normal one as to be confused. Thus to execute similar functions to the human ones to the point of making it possible to confuse them, they would have to possess such types of reflections, which implies consciousness. And he mocks: "Zimboes thinkz

[32] This argument is based on the Necessity Postulate: if something is identical, it is necessarily identical in all possible worlds. If there is a zombie-world, states of consciousness are not identical to brain states. Thus physicalism is false.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

they are conscious, thinkz they have qualia, thinkz they suffer pains – they are just 'wrong' (…)" (*Idem*, p. 173). But how would we know we are not wrong too?

Dennett offers us an analogy to show how this idea is inconceivable: he tells us to imagine removing health and still leaving all of our body functions intact. Such idea is not conceivable, so how can the p-zombies be? It means the difference between zombies and conscious beings can only be illusory. This philosopher compares the idea of p-zombies with the idea of "epiphenomenal gremlins": both of these ideas are equally "silly".

Another philosopher who wants to demonstrate that p-zombies are not conceivable is Eric Marcus. In his article "Why Zombies Are Inconceivable" the author seeks to contradict this thought experiment and its implications using Chalmers' terminology.

To do so he challenges us to imagine how it is like to be Abe Lincoln. Imagine Abe Lincoln has a zombie twin, "Zombie-Abe". We can imagine how it's like to be Abe Lincoln on the first-person, subjectively, employing the Cartesian language in which usually Chalmers incurs. Still we cannot imagine "Zombie-Abe". We can imagine an empty Abe in the third-person, a "Zombie-Abe" just like we can imagine empty spaces. But here Chalmers is asking us to imagine the lack of consciousness. In comparison, we could be tempted to think that the lack of consciousness is possible to imagine just as it is possible to imagine the lack, for example, of pains. However, it is not a good analogy given the fact that we experience lack of pain very often, and the same can't be said about consciousness. We simply do not experience anything beyond the things we are conscious of; if the opposite happened it would not have the name of conscious experience. Therefore, to imagine zombies is actually not to imagine anything.

In this case, not to imagine anything or imagine total lack of subjective experience is, for Marcus, an impassible bridge, even more impassible than imagine, ironically, how it is like to be a bat. At this point Marcus suggests that trying to imagine zombies we would confuse not imagine something (not imagine conscious states) with imagining nothing.

With Chalmers taxonomy we can now understand how, even if prima-facie conceivable, this hypothesis is ideally inconceivable; and being inconceivable that a zombie world is real, it is also secondarily inconceivable. If we accept Dennett's thesis that this idea is auto-contradictory we can affirm that it is not also negatively conceivable; yet, according to Marcus, even if we don't say there's an *a priori* contradiction, in fact zombie worlds are positively inconceivable for we cannot imagine coherently the real existence of zombies.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Even if we couldn't find an *a priori* contradiction in imagining zombies, in fact zombie worlds are positively inconceivable for we cannot imagine coherently real existence of zombies. In this way, the philosopher considers he has brought down the path Chalmers made as the most adequate to the possibility of philosophical zombies.

So it is possible that the p-zombies argument doesn't succeed in refuting the functionalism and a computational theory of mind. But even if it did, a refutation wouldn't threaten these theories. That happens given a division Chalmers does before Dennett's critics in his paper about Singularity. In "The mystery of David Chalmers", Dennett says it is strange how on one hand Chalmers vigorously defends a computionalist theory of mind with a functionalist assumption, but, on other hand, has those visions about consciousness as an intrinsic and fundamental property, because such an assumption would imply Chalmers being a type-A materialist or a functionalist, theories he criticizes. Dennett questions then "Why is Chalmers not a type-A materialist? He gives very good arguments for type-A materialism, and finds no flaws in them" (Dennett, 2012, p. 89).

However, Chalmers affirms Dennett is not taking the distinction he makes between functionalisms in two different subjects into account.

(S1): the question of the *relation* between consciousness and physical correlates;
(S2): the question of knowing whether the physical correlates of consciousness *are biological or functional.*

On the one hand, in relation to (S1), Chalmers criticizes the functional theory through his modal arguments, such as the p-zombies argument. The hypothesis of the existence of p-zombies isn't, however, to be considered as a plausible hypothesis in our actual world.

Thus, on the other hand, in relation to (S2), Chalmers defends the functional position present in the premise (16). To (S1) modal arguments are used, such as the p-zombies one. To (S2) Chalmers uses an argument exposed in his paper "Absent Qualia, Fading Qualia, Dancing Qualia", in which he refers to the notion of "Organizational Invariance", defined as follows:

> In general, if a property is not an organizational invariant, we should not expect it to be preserved in a computer simulation (a simulated rainstorm is not wet). But if a property is an organizational invariant, we should expect it to be preserved in a computer simulation (a simulated computer is a computer). (Chalmers, 2010, p. 40)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Chalmers considers consciousness[33] a property of this kind and, this way, he defends that any two systems with the same organization will have the same conscious experience.

Thus we finally understand the functionalist assumption present in premise (16) resulting in the philosopher's optimism relatively to the mind-uploading possibility.

3.2. The Chinese Room Argument

The argument and thought experiment commonly known as “Chinese Room” was published in 1980 by the American philosopher John Searle. The Chinese room thought experiment describes someone who, not knowing how to speak Chinese, is closed inside a room where there are Chinese symbols inside boxes; this person has a book of instructions in English which explains how to combine the Chinese symbols and how to send sequences of Chinese symbols out the room, when other Chinese symbols are introduced in the room, through a small opening, so the person outside doesn't realize what is going on inside the room; the person inside doesn’t know that the introduced symbols are “questions” and the ones that go out are “answers”. We can conclude that the system, in its whole, talks and understands Chinese, in the perspective of the people who are outside. Thus, the system passes the Turing Test, even if the person inside knows she doesn't understand Chinese at all. Searle concludes that this thought experiment shows the possibility of a system which has an “atributed intentionality” but not “intrinsic intentionality” (genuine semantics), that is, that syntax is not enough, but the comprehension of involved symbols is also necessary, taking down then the project of Strong AI (cf. Searle, 1984, pp. 29-32).

You can confront the illustration of the experiment in Fig. 8-3.

[33] However, as we'll see next, a functionalist of a biologist type will not accept this idea of the existence of properties which are independent of a particular substract. (cf. Pigliucci, 2014, p. 122)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

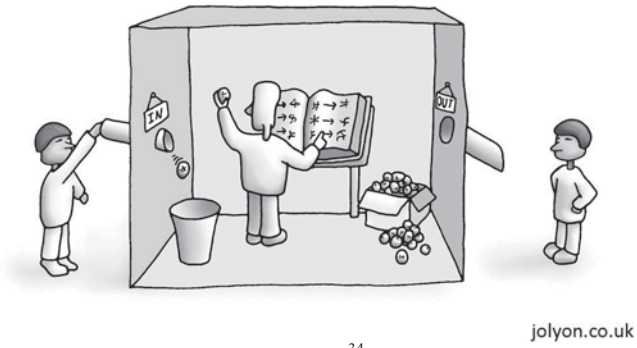

Fig. 8-3. representation of the Chinese Room[34]

As we can already tell, this argument tries to demonstrate that any functionalist notion that is based in the type of digital computation is deeply wrong. We can formulate the argument this way:

(20) If AI is true, then any system running the "understand Chinese" program really understands Chinese.
(21) The person can run the "understanding Chinese" program without in fact understanding Chinese.
(22) There is at least one system which runs the "understand Chinese" program without in fact understanding Chinese.
(23) Therefore, AI is false (follows from 20 and 22). (cf. Mandik, 2014, p. 98)

Adapting, then, the Chinese Room Argument above presented, the argument works well against the specific "Turing Machine" type of functionalism because both can be affirmed in terms of their programs. See: this kind of functionalism has in its base programs because it's compromised with the idea that mental states can be defined in terms of an instruction program from a look-up table to a Turing machine. But at the base of the Chinese Room Argument is also an affirmation that a "Chinese understanding" program can be executed by a person who doesn't really understand Chinese, undermining definitely the described project: we can't affirm both at the same time without contradiction.

Is this a good argument? This argument was scrutinized in detail for too much time in the philosophical literature.[35] As such, we'll briefly

[34] Cf. https://www.emaze.com/@ALFLCWIQ/a.i
[35] For example, Jerry Fodor (1980) conforms to Searle on the importance of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

present an answer which seems to be effective against this thought experiment. But firstly we ask you to pay attention to Fig. 8-4:

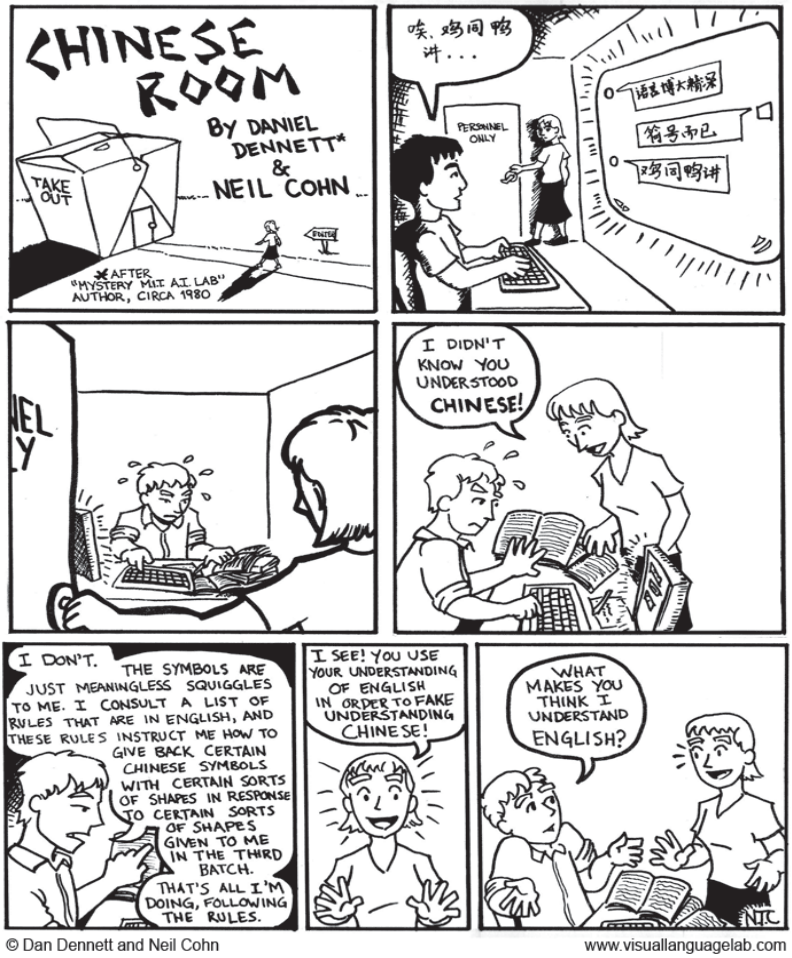

Fig. 8-4. critic of the Chinese Room [36]

semantics. However, unlike Searle, Fodor thinks that the formal symbols of a computer may have semantic properties (see Semantic Theory presented in Fodor (1990)).

[36] Image reference: http://www.visuallanguagelab.com/chinese_room/index.html

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The idea[37] is to show that "understanding" is in the total system, not a part of this system. That is, this answer accepts that the person inside the Chinese room does not understand Chinese (being only the CPU of the system), but that the entire system displays all the conditions of understanding Chinese.

However, in trying to answer this objection, Searle asks us to internalize all parts of the system in the person's brain. We memorize the data of experience and symbols, and so we build a Chinese room in our head. But such a strategy seems to worsen the situation: first, it is impossible for a brain to memorize so much information at the same time; but even assuming that it was possible, then this brain would have reached such a level of complexity that we could conclude that the person, after all, really understood Chinese.

Updating this argument to the subject today, Searle would say that Kurzweil's estimates are simply wrong because his estimate of the computational complexity of the brain is incorrect: this approach assumes that all information processing in the brain can be represented by a combination of pulses when neurons fire (action potentials) and the number of synapse receptors each neuron possesses. After that it would be enough to multiply the estimate by the number of neurons and their synapses and add everything else. Finally, multiplying the percentage would trigger a neuron which is at 200 actions potentials per second. This model was consistent for some years in neuroscience. However, we now know it is quite wrong and leaves out many of today's discoveries[38] (for example, the phenomenon of "Backpropagation", i.e., signals traveling from the sum to the dendrites; the action potential is reflected within the axons and travels in reverse (only this information could double the system's complexity).

Despite that the argument seems not to have enough strength, Searle wants to draw the attention to the fact that the brain is an element constructed by a natural and biological evolution and that only this type of body can have properties of higher order as consciousness and intelligence. Note then that what Searle would say in this matter would be something like: it is not enough to reproduce the parts of the system (anti-functionalist thesis), but also the material or medium that this system is

[37] Besides, we can use the same structure argument to show that the human brain itself has no understanding of any language – so we proved that we are all a Chinese room!

[38] For a detailed look of particular errors of the estimation approach of Kurzweill, cf. Dettmers, 2015.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

made of is relevant to the creation of a conscious system. This is in the idea presented in Conclusion 4 presented in his famous book "Mind, Brain and Science":

> For any artefact that we might build which had mental states equivalent to human mental states, the implementation of a computer program would not by itself be sufficient. Rather the artefact would have to have powers equivalent to the powers of the human brain. (Searle, 1984, p. 41)

For Searle what matters is the causal power[39] that the system has. Thus, the problem is not if a silicon chip can fully replace a neuron. But silicone seems to have nothing to do with computation, in a sense of being a formal and symbolic abstraction that we can implement in any medium we want. For Searle, it is a problem of empirical order if silicon has the same causal power that a human brain has – as an argument is not a priori – although Searle thinks that no medium can have this causal power for its biological precondition that is influenced by the theory of evolution. This is a factual position of Searle. But the philosophical position of Searle is that having only the symbolic form itself is not sufficient to ensure the presence of consciousness.

Thus, Searle can be seen as a proponent of a functionalist theory of a biologist type, in which it is important not only to be right in the reunion of the system parts, but also the chosen material (whether organic or inorganic). In addition, Searle is against any kind of purely functionalist Computational Theory because it does not respect the condition of causal power required.

3.3. Analysis of the Summary of the Arguments

We can summarize the analysis in 3.1. and 3.2. with the following theses:

T1: The composition and arrangement of parts *x* is sufficient to produce all *y* properties. (conclusion of part 3.1.)

[39]Another quote of Searle that goes in the same direction: "Part of the point of the present argument [Chinese Room] is that only something that had those causal powers could have that intentionality. Perhaps other physical and chemical processes could produce exactly these effects; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. That is an empirical question (…)." (Searle, 1980)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

T2: The material (causal power) of the parts of *x* is required to reproduce the *y* properties. (conclusion of part 3.2.)

Now, on the one hand, a functionalist of a computational type (like Chalmers) only accepts T1 as true. Therefore, we can call them Weak Funcionalists. The problem of consciousness is formulated only as a programming problem. On the other hand, a functionalist of a biologist type (such as Searle) requires that T1 and T2 are simultaneously true. Therefore, we can call them Strong Funcionalists.

But both seem to have two opposing positions on the problem of the mind-uploading: a Weak Funcionalist seems to have no problem in assuming its possibility, since we could appropriate computational theory – he is then optimistic. A Strong Funcionalist can be against that possibility, assuming only a biologically evolved process could be conscious – he is then pessimistic.

After this separation, is there any way to join these two positions? We will conclude with a possible hypothesis that seeks to unite the two presented intuitions: through a simulation of the evolutionary process itself.

4. A Possible Way

An idea we think might be promising in a combination of functionalisms, the strong and the weak one, is trying not to simulate consciousness but rather simulate the process that led to the emergence of consciousness in organic matter (in this case, animals), applying it to non-organic matters.

Thus both "functionalisms" will be satisfied. We know, through the studies of sciences, the mechanisms of the evolutionary process have made the emergence of consciousness possible through a fairly long period of time. It is possible today to artificially computationally simulate this process in a short period of time.

This type of computing inspired in biology and evolution uses techniques precisely inspired on these fields relative to the fact that consciousness exists in a body which behaves in a particular environment and learns interacting with it, conversely to the developments of a more “traditional” AI field.[40]

[40] “[Typical AI] ended up neglecting fundamental aspects of biological intelligence, such as physical embodiment, behavioral autonomy, self-healing,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

An example of a project that applies this type of computing is the “Biota.org", which studies the natural and artificial systems, and inserts "digital biota" or "cyber biota" in robots. Digital Biota is a kind of autonomous software that self-replicates and is embodied in viruses, genetic algorithms and general adaptive networks. The objects of the software interact with their environment, they are able to multiply, learn and change, being affected by "natural selection” whose rules have been programmed to evolve.

They are then programmed evolutionary algorithms that use models such as the aforementioned genetic algorithms. Such algorithms have structures that evolve according to “selection", as mutations and recombinations to adapt. For example, some functions use measures of adaptation to probabilistically select individuals who best suit to their environments – individuals go through a trial and error, learning, changing and evolving. Josh Bongard, in his Morphology Laboratory, Evolution and Cognition at the University of Vermont, uses the "Artificial Ontogeny" – he creates virtual eggs that grow into adult robots, finding what the most suitable robots are to particular tasks through computational simulations of these robots in certain environments, as you can see in Figs. 8-5. and 8-6.:

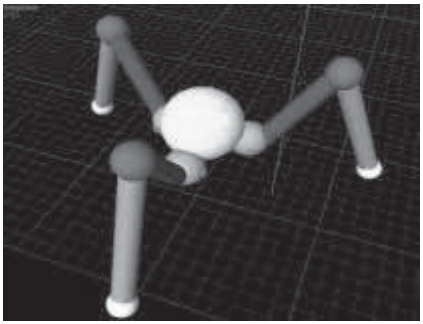

Fig. 8-5. artificial ontogeny I

social interaction, evolution and learning, that make biological organisms prone to errors and sometimes difficult to predict, but also so successful to survive in unknown and changing environments.” (Floreano & Mattiussi, 2008)

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

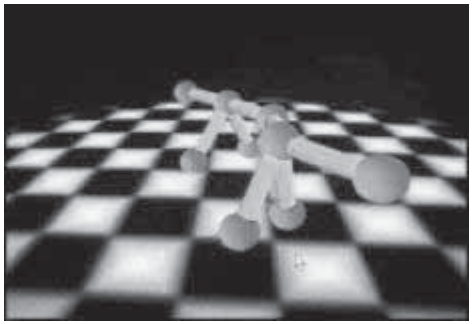

Fig. 8-6. artificial ontogeny II

If so, through this type of research, robots evolve through a selection process, acquiring the same causal powers that are in the minds found in organic substrates, satisfying (T2), i.e. the evolution that allowed organic systems to become conscious will allow non-organic systems to do it too. The existence of robots with minds facilitates the transfer process of my mind to the robot. To this end, the system which produces the mind would only have to be arranged in the same way as the system (the body and the brain) which produces my mind, being then satisfied (T1).

So, with the help of Chalmers and Searle insights, and new promising innovations from the field of new Artificial Intelligence, the mind-uploading may become a possibility in the near future.

Bibliography

Block, N. (1981), "Psychologism and behaviorism", in *Philosophical Review* 90 pp. 5-43.

Bradford, S. (2012), "A Framework for Approaches to Transfer of a Mind's Substrate", in *International Journal of Machine Consciousness*, Vol. 4, No.1, pp. 23-34.

Chalmers, D. (2014), "Uploading: a Philosophical Analysis", in *Inteligence Unbound: the future of uploaded and machine minds* (ed. Russell Blackford and Damien Broderick), Oxford: Wiley Blackwell.

—. (2012), "The Singularity: a Reply to Commentators", in *Journal of Consciousness Studies* (7-8) pp. 141-167.

—. (2010), "The Singularity: a Philosophical Analysis", in *Journal of Consciousness Studies* 17:7-65.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. (2002), "Does Conceivability Entail Possibility?", in *Conceivability and Possibility* (T. Gendler & J. Hawthorne [eds.]), Oxford: Oxford University Press.
—. (1996), *The Conscious Mind*, Oxford University Press.
—. (1995), "Absent Qualia, Fading Qualia, Dancing Qualia", *in Conscious Experience*, Thomas Metzinger (ed.), Paderborn: Ferdinand Schöningh, pp. 309-328.
Dennett, D. (2012), "The Mystery of David Chalmers" in *Journal of Consciousness Studies,* 19, Nos. 1-2, pp. 86-95.
—. (1998), *Brainchildren, Essays on Designing Minds*, Cambridge, MA: MIT Press and Penguin.
—. (1995), "The Unimagined Proposterousness of Zombies", *in Journal of Consciousness Studies* 2 (4):322-26.
—. (1991), *Consciousness Explained*, Boston, MA: Little, Brown and Company.
Dreyfus, H. (1972), *What Computers Can't Do*, New York: MIT Press.
Floreano, D. & Mattiussi, C. (2008), *Bio-Inspired Artificial Intelligence: Theories, Methods and Technologies*, Cambridge: MIT Press.
Fodor, J. (2000), *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.
—. (1990), *A Theory of Content and Other Essays,* Cambridge, MA: MIT Press.
—. (1981), "The mind-body problem", in *Scientific American* 244:114-25.
—. (1980), "Searle on what only brains can do", in *Behavioral and Brain Sciences* 3: 431-2.
—. (1975), *The Language of Thought*, New York: Thomas Crowell.
Horst, S. (2009), The Computational Theory of Mind, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/computational-mind (accessed September 5, 2015).
Koene, R. (2014), "Feasible Mind Uploading", in *Inteligence Unbound: the future of uploaded and machine minds* (ed. Russell Blackford and Damien Broderick), Oxford: Wiley Blackwell.
Kurweil, R. (2006), *The Singularity is Near*, New York: Viking Penguin.
—. (1999), *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York, NY: Penguin Books.
Lucas, J. R. (1961), "Minds, machines, and Godel"., in *Philosophy* 36: 112-27.
Mandik, P. (2014), *This is Philosphy of Mind*, Oxford: Wiley-Blackwell.
Marcus, E. (2004), "Why zombies are inconceivable" in *Australian Journal of Philosophy*, 82 (3): 477-90.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

McCulloch, W.S. & Pitts, W. H. (1943), "A Logical Calculus of the Ideas Immanent in Nervous Activity'", *in Bulletin of Mathematical Biophysics* 7, 115–133.

Nicolelis, M. & Cicurel, R. (2015), *The Relativistic Brain*, São Paulo: Kios Press.

Penorse, R. (1994), *Shadows of the Mind*, Oxford University Press.

Pigliucci, M. (2014), "Mind Uploading: a Philosophical Counter-Analysis", in *Inteligence Unbound: the future of uploaded and machine minds* (ed. Russell Blackford and Damien Broderick), Oxford: Wiley Blackwell.

Putnam, H. (1967), "The Nature of Mental States", in Art, Mind and Religion (eds. W.H. Capitan & D.D. Merrill), Pittsburgh University Press 1-223.

—. (1960), "Mind and Machines", in *Dimensions of Mind* (ed. S. Hook), New York: Pocket Books.

Ravenscroft, I. (2005), *Philosophy of mind: a beginner's Guide*, New York: Oxford University Press.

Sandberg, A. & Bostrom, N. (2008), *Whole Brain Emulation: A Roadmap*, Technical Report #2008-3, Future of Humanity Institute, Oxford University.

Searle, J. (1984), *Minds, Brains and Science*, Cambridge, MA: Harvard University Press.

—. (1980), "Minds, Brains and Programs", in *Behavioral and Brain Sciences* 3 (3): 417–457.

Teixeira, J. de F. (2004), *Filosofia e Ciência Cognitiva*, Editora Vozes.

Tipler, F. (2012), "Inevitable existence and inevitable goodness of the singularity", in *Journal of Consciousness Studies*, 19: 183-193.

Websites

Hendricks, M. (2015, September 15), *The False Science of Cryonics*, Retrieved December 26, 2015, from https://www.technologyreview.com/s/541311/the-false-science-of-cryonics/

Neal, M. (2013, August 7), *For One Second, a Supercomputer Mimicked the Human Brain,* Retrieved September 27, 2015, from http://motherboard.vice.com/blog/for-one-second-a-supercomputer-mimicked-the-human-brain?utm_source=mbfbvn.

Dettmers, T. (2015, July 27), *The Brain vs Deep Learning Part I: Computational Complexity — Or Why the Singularity Is Nowhere Near*, Retrieved October 12, 2015, from

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

http://timdettmers.com/2015/07/27/brain-vs-deep-learning-singularity.

Youtube videos

Wall Street Journal (2015, January 28), *Scientists Upload Worm's Mind Into a Lego Robot,* Retrived from https://www.youtube.com/watch?v=2_i1NKPzbjM.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER NINE

GODSEED: BENEVOLENT OR MALEVOLENT?

ERAY ÖZKURAL

It is hypothesized by some thinkers that benign looking AI objectives may result in powerful AI drives that may pose an existential risk to human society. We analyze this scenario and find the underlying assumptions to be unlikely, as well as the premises of the argument. We argue that the AI eschatology stance is not scientifically plausible: more intelligence helps avoiding accidents and learning about ethics; and we also argue for the rights of brain simulations. We may still conceive of logical use cases for autonomy. We examine the alternative scenario of what happens when universal goals that are not human-centric are used for designing AI agents. We follow a design approach that tries to exclude malevolent motivations from AI agents; however, we see that objectives that seem benevolent may pose significant risk. We consider the following meta-rules: preserve and pervade life and culture, maximize the number of free minds, maximize intelligence, maximize wisdom, maximize energy production, behave like human, seek pleasure, accelerate evolution, survive, maximize control, and maximize capital. We also discuss various solution approaches for benevolent behavior including selfless goals, hybrid designs, Darwinism, universal constraints, semi-autonomy, and generalization of robot laws. A "prime directive" for AI may help in formulating an encompassing constraint for avoiding malicious behavior. We hypothesize that social instincts for autonomous robots may be effective such as attachment learning. We mention multiple beneficial scenarios for an advanced semi-autonomous AGI agent in the near future including space exploration, automation of industries, state functions, and cities. We conclude that a beneficial AI agent with intelligence beyond human-level is possible and has many practical use cases.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

1. Introduction

An interesting question about AGI (artificial general intelligence) agent design is how one would build an "angelic" autonomous AGI agent. Would it be possible to make some kind of *angel's* mind that, by design, achieves only good? Philosophically speaking, is there any cosmic standard of ethics (since *angel* is just a mythological fantasy)? In this paper, we would like to define universally benevolent AGI objectives, also discussing what we consider to be malevolent objectives, as well as the limitations and risks of the objectives that we present.

This is also a common question that many seek a somewhat easier answer to in the form of "friendly AI" which has been explained in [12]. In that paper, Yudkowsky defines friendly AI very generally as a superintelligent system that realizes a positive outcome, and he argues laboriously that abandoning human values will result in futures that are worthless from a human point of view, and thus recommends researchers to seek complex value systems (of humans) for embedding in AI's. While that is a challenging goal in itself, we think that the alternatives have not been exhaustively researched. One idea that comes to mind is that some of the better aspects of humanity may be generalized and put into a universal form that any intelligent, civilized agent, including extraterrestrials, will agree with. Furthermore, the friendly AI approaches (putting human desires at the forefront) may have some shortcomings in my opinion, the most obvious is that it places too much faith in humanity. They seem also ethically ambiguous or too anthropocentric, with such assumptions that machines would be considered "beneficial" if they served human desires, or that they would be deemed "good" if they followed simple utilitarian formulations which seem to try to reduce ethics to low-level properties of the human nervous system. First, it has not been persuasively explained what their utility *should* be. If for instance positive utilitarianism were supposed, it would be sufficient to make humans happy. If human society degenerated as a whole, would this mean that all resources would be spent on petty pursuits? If a coherent extrapolated volition [11] were realized with an AGI agent, would this set our sights on exploring other star systems, or spending our resources on such unessential trivialities as luxury homes and sports cars? Would the humans at one point feel that they have had enough and order the AGI to dismantle itself? The human society is governed mostly by the irrational instincts of apes trapped in a complex technological life, and unfortunately not always with clear goals; will it ever be possible to refine our culture so that only significant ideas take the lead? That sounds more like a debate of social theory than AGI

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

design. Or suppose that there are AGI agents that have become powerful persons and are friendly to humans. Such subservience would be quickly exploited by the power hungry and corrupt humans. Then, would this not lead to unnecessary conflicts, the oppression of the greedy and the rule of the few over the many, unless many other social changes are enforced? Or should we simply wish that social evolution will necessarily bring the best of us?

I do not think that the present subject is a matter of technical debate, thus I will approach the subject philosophically, from a bird's eye view at 10000 feet. If we did not design the AGI agent around anthropocentric concepts like human-friendliness, as if agents are supposed to be exceptionally well behaving pets, would it be possible to equip them with motivations that are universally useful/benevolent, applicable to their interactions with any species, intelligent machines and physical resources? Would it be possible to grant them a personal existence far beyond us, with motivations that far exceed ours? What would they do in a remote star system when they are all alone by themselves? What kind of motivations would result in occasional "bad" behaviors, and what are some of the universal motivations that we may think at all? Another important question is how much potential risk each such AGI objective/motivation presents to us. I shall try to answer questions such as these in the present article.

2. Misprogrammed AI Agents Do not Pose an "Existential Risk"

AI eschatologists believe that a misprogrammed AI agent can destroy the world with a significant probability. AI eschatology literature mainly blows the conclusions of Omohundro's philosophical article [5] out of proportion, which argues for AI drives that will result from specifying a benign looking goal, such as maximizing paperclips in the world. Surely, such an objective must involve turning all matter to paperclips, hence it should destroy the world in order to achieve that goal, the argument goes. Besides the obvious bravado of the said argument, it is also ridden with a typical fallacy of making an improbable event seem probable. A long chain of weak causes (and strong assumptions) usually results in an inference with very low probability; beneath a certain level we are forced to regard it as improbable, such as Bertrand Russell's notorious earth-orbiting lovely ceramic teapot. Bostrom and Yudkowsky repeatedly ask us to concede to a long chain of unlikely events, the conjunction of which will result in the eradication of our species. As a "solution", they often

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

mention building a UN controlled “friendly AI” that will prevent others from building such destructive “demonic intellects”. Let us start with unveiling their tacit assumptions, showing the improbability of any such risk.

AI must be an agent: That is quite untrue. A kind of AGI program the author is working on is completely “passive" and not an agent at all, yet has all the intelligence that an agent can have. At any rate, most AI programs are not agents, the most useful kind is machine learning applications like speech/face recognition.

AI agents must be autonomous: No, AI agents do not need to be fully autonomous. They would rather be programmed to do whatever task is needed. It is a quite silly idea to have to convince a robot to do a job, and that is not how it should be. To replace labor, we must use AI in the most effective way; emulating a person is certainly not necessary or desirable for this kind of application. This also seems like an unlikely, arbitrary assumption that is based on a confusion that the AIXI model is the only way to formulate an AGI system. AIXI is a reinforcement learning model, it models a general kind of utility-optimization agent, but it is not necessary to make an autonomous agent to build intelligence into an application.

Even a question/answer machine is dangerous: No, it is not. A Q/A machine is completely "passive", it only learns and solves problems posed. It has no will of its own, and has no goals whatsoever, apart from giving the correct answer to a problem, which constitutes pure intelligence. A typical example of a Q/A machine is a machine learning classification problem, such as telling apart whether a mushroom is edible or poisonous based on its attributes. The way they thought this would be dangerous is: a politician comes and asks "What must I do to win this election?" and then the machine tells him to do all kinds of sinister things ending humanity. Of course, that is a ridiculous and implausible science fiction scenario that is not worth elaborating.

AI will necessarily have harmful AI drives: Omohundro in his paper argued that pursuing an innocent looking objective like "maximizing the number of paperclips" could have harmful consequences, since the AI agent would do anything to reach that objective. It would also have animal-like drives, such as survival. Omohundro’s analysis does not apply to any kind of design and motivation system. Autonomous robots with

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

beneficial goal systems have been discussed by Ben Goertzel [1]. I have offered a conceptual solution to designing motivation systems: open-ended and selfish meta-goals are harmful to some when applied to fully autonomous agents, but there are many ways to fix this, such as removing full autonomy from the system, adding universal constraints (such as non-interference, advanced "robot laws", i.e., legal, logical AI agent), and making closed-ended, selfless motivations, as will be discussed in the present paper. The simplest solution, however, is to avoid autonomy in the first place, as well as goals that are animal-like (such as maximizing pleasure).

Human preferences may be made coherent: They contradict wildly and manifestly. The views of superstitious folk, in majority, contradict with those of intelligent people. It is hard to see who would be fit to train such an agent even if we picked preferentially. The sad story is that humans in general are not good at ethics and they have many wrong and harmful ideas about the human society, and training from the world at large would only be worse.

A UN controlled AI dictatorship is plausible: It is neither plausible nor desirable. It is diametrically opposed to democracy and freedom. Banning AI research is essentially banning all computer research. AI is just an apex of computer science. When one bans AI, they have to also ban computer science. That is how absurd that view is, it is even less plausible than regulating cryptographic software. On the other hand, no person would want to give up his sovereignty to an AI controlled by UN. It is also completely unreasonable since most communities demand decentralized and democratic governance.

Singularity can occur anywhere: It cannot. It is doubtful whether a "singularity" will occur. More likely, a higher technological plateau will develop, no real or approximate singularity will occur because there are physical bottlenecks that will cause very significant slowdowns after 2030. However, even if we assumed there were no bottlenecks (and according to my projections that would mean a singularity by 2035 [8]), the theory concerns the whole globe, not a small subset of it. A rapid technological evolution can only be funded by a very large nation at the very minimum, and even then it would be very unlikely. The likely event is that the whole globe will participate in computer technology, as it has in the past. It is pseudo-science to think that it can happen in a garage or even by a single nation or megacorporation. In reality, so-called infinity point, or

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

singularity is quite unlikely to happen for physical processes such as required experiments and manufacturing form a serious bottleneck. In all likelihood, we will build computers much faster than a human brain, but that will still take many decades, and we will not reach physical limits of computation any time soon, because that would require us to form extreme physical regimes we are not capable of yet.

Goertzel reviews the problems in the AI eschatology folklore in a lucid paper that distills the problem with the eschatological stance to its essence: that it is an informal rather than a scientific argument [2]. We should further emphasize that there is no real evidence about the probabilities claimed; to obtain a high probability like 20% for a human extinction event we would have to be assigning a quite high probability to this supposed misprogrammed AI monster that breaks out of the lab and kills all humans. We may also assign very low arbitrary probabilities to individual conditions that make up their argument, which Goertzel clarifies in his blog as:

1. If one pulled a random mind from the space of all possible minds, the odds of it being friendly to humans (as opposed to, e.g., utterly ignoring us, and being willing to repurpose our molecules for its own ends) are very low.
2. Human value is fragile as well as complex, so if you create an AGI with a roughly-human-like value system, then this may not be good enough, and it is likely to rapidly diverge into something with little or no respect for human values.
3. "Hard takeoffs" (in which AGIs recursively self-improve and massively increase their intelligence) are fairly likely once AGI reaches a certain level of intelligence; and humans will have little hope of stopping these events.
4. A hard takeoff, unless it starts from an AGI designed in a "provably Friendly" way, is highly likely to lead to an AGI system that doesn't respect the rights of humans to exist.

These are all scientifically implausible speculations that have no real counterpart in either philosophy of ethics or AI literature. By making every step of their argument only slightly fantastical, they succeed in reaching a fantasy land that is quite incredible. The first assumption we may term as "Intelligence is the original sin" doctrine. It may sound reasonable until one considers that we have not designed a single human-level intelligent agent besides our own. We only know of animals, that are quite similar to our own architecture. We have not made a comprehensive

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

exploration of the whole space of possible mind designs yet. Therefore, we simply do not know if intelligence begets evil as scholastic philosophers might have agreed to. The second is also speculative, both philosophically and technically: if human values are fragile, then how can we depend on them in any way? A human may shape his behavioral patterns in many ways, attaining many cognitive and behavioral characteristics as his default mode of operation, including ethical ones, such as being violent, or harmful. It is premature to assume more intelligence does not and cannot help an agent improve its ethical knowledge. AI theory suggests that it should be able to. Then, why assume such divergence is possible? That seems like a textual confusion that confounds AI eschatologists. However, in the world of actual intelligent agents, we see that more intelligence helps agents understand the world better, including ethics, and formulate better goals and plans. It is misleading to think that assigning a ridiculous goal like maximizing paperclips, with obviously harmful consequences, is a good example of intelligent agent design. For intelligent action requires intelligent goals, which we can program as present article suggests. We can also build as many constraints as we like *into the design*, requiring no insane "countermeasures" like kill-switches, that AI eschatologists are fond of. The improbability of the hard takeoff idea has already been explained, but to reiterate, the infinity point hypothesis is an abstract macro-economic model that is only talking about a supposed extrapolation of Moore's law; it is not going to happen in that exact way, it will be much slower and require the co-operation of the entire globe. I will attempt to propose a more realistic model of technological evolution in future work, nevertheless, those constitute the Achilles' heel of the AI eschatology argument. Even if a random mind would be evil, which sounds like a fantastical notion, there will not be a hard take-off, and in particular a single agent will not achieve it. These are so improbable events that it is hard to assign a probability to them, but try as we might, we would have to say that the conspiracy theories that extra-terrestrial intelligences are governing the world are much more probable than the hard take-off assumption. Such extraordinary claims do require extraordinary evidence as Carl Sagan would say, and there is no such evidence for the hard take-off claim, or any of the conjunctive assertions here, which leaves the conjunctive argument itself highly improbable, not truly worthy of our consideration.

Of course, robots can be dangerous. In accidents, heavy industrial robots have already killed people. Increasing their intelligence could certainly help prevent accidents, which was covered in Asimov's robot laws. Only high intelligence could react rightly to an accident and save a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

person's life in time. Therefore, if robots are to be abundant, we do need more advanced intelligence to prevent harm to humans. However, that does not mean at all that the robot must be human-like in personality or in cognitive architecture. Briefly, it does not need to be a person. I call this the "anthropomorphic AI fallacy", and I note that it is widespread. A machine can be much more intelligent than a human, yet may entirely lack any human-like personality or autonomy. In fact, the most practical use of AGI software would be through very deep brain-machine-interfaces, which would communicate our questions and receive answers rapidly. In robotics, this would happen, as translating our goals to robotic devices, or remote controlling them intelligently.

Should we grant personhood to intelligent, autonomous robots? We should, at least to a certain kind of robot: a robot equipped with a brain simulation. The digital person-branch of a biological person will already know and understand human conventions, and will be responsible for his actions. And that is the only way to have practical technological immortality; if my immortal, technological form did not have any rights, what would the point of its existence be? It is our cyber progeny that will colonize the solar system and exoplanets, and thus we will have to concede rights to our progeny. I would certainly not allow my brain simulation to be equipped with a killswitch as Bostrom demands.

Likewise, for autonomous agents, we may envision a system where there are rigid laws controlling their behavior; I thus prefer Mark Waser's libertarian solution to this problem of AI ethics. However, I must underline that we cannot assume any AI agent will be responsible for its behavior, before we make sure that it has the capability and the right cognitive architecture. Both Steve Omohundro and I accept that we may program inane motivations that would turn out to be harmful, however, just as a human can have a somewhat stable psychology, so can a robot. We can allow such artificial persons – like Commander Data in Star Trek, which is much better science fiction than AI eschatology – if and only if we are certain of its psychological qualities, it is true that we must not hurry with such projects.

Would not it be horrible that robots were used for crimes? Indeed, robots are already being used for horrible war crimes. Drone strikes are commonplace, and few raise an eyebrow over that, instead gleefully cheering the onset of the combat robotics. In the future, most wars will be fought by machines, and these machines do not need any more than rudimentary intelligence. Most high-tech weaponry are robots, such as a guiding missile. In the future, most will be robotic. Thus, perhaps, we should question the ethics of our fellow, naturally not-so-intelligent

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

humans, rather than extremely intelligent, autonomous robots that do not exist.

That technology can be used to inflict harm is not a good enough reason to ban it, because the benefits often outweigh the harms. For AI, many orders of magnitude so. People must instead be worried about people who will use robots for their evil deeds. On the other hand, AI technology will be pervasive, it will change the very way we use computers. Computers could not really create much useful information on their own before, we mostly created and edited data on them. Now, computers will create useful data on their own. AI is not just some robotics technology, it is a wholly new era of computing. Even the capability to understand and react to human language will vastly change the computing landscape.

3. Is the Concept of Malevolence Universal?

Previously, Omohundro identified basic AI drives in reinforcement learning agents with open ended benign looking AI objectives [5]. In the end, when we share the same physical resources with such an agent, even if the initial intention of the utility programming was benign, there will be conflict, especially in the longer run, and harm may come to humans. I will in this article instead ask if there are benevolent looking universal objectives, and whether there might be any risk from assuming such objectives in an AI agent.

Let us thus consider what is ever evil. I suspect, intuitively, that a prior source of many evil acts is selfish thinking, which neglects the rest of the world. Being selfish is not only considered evil (traditionally) but it defies rationality as well, for those species that may collaborate are superior to any single individual. There is however much disagreement about what is evil, so I will instead prefer the more legally grounded term of malice or malevolent acts. In a galactic society, we would expect species to collaborate; if they could not trust one another, then they would not be able to achieve as much. Another example is science: science itself is a super-mind which is an organization of individuals, working in parallel, in civilized co-operation and competition, so it too requires a principle of charity at work. When that fails, the public may be misinformed.

Here are some examples of malevolent acts: if someone disrupted the operation of science, if someone gave you misinformation on purpose, if someone misappropriated resources that would be much beneficial for the survival and well-being of others, if someone tried to control your thoughts and actions for his advantage, if someone destroyed life and

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

information for gain, if someone were indifferent to your suffering or demise. Thus, perhaps biologically, malevolent behavior goes back to the dawn of evolution when symbiotic and parasitic behaviors first evolved. However, the most common feature of malevolence is a respect for self foremost, even when the malevolent one seeks no selfish reward. Then, perhaps I cannot assure a perfectly “angelic” agent, for no such thing truly exists, but I may at least design one that lacks a few common motivations of many acts that we consider malevolent. See [10] for a similar alternative approach to universal benevolence.

In theory, an obvious approach to avoid malevolent acts would be to try to design a "selfless" utility function, i.e., one that maintains the benefit of the whole world instead of the individual. This criterion will be discussed after some AI objectives have been presented. Other important questions were considered as well. Such an AI must be economically-aware, it must lean towards fair allocation of resources, instead of selfish (and globally suboptimal) resource allocation strategies. A scientific instinct could be useful, as it would go about preserving and producing information. It might have an instinct to “love” life and culture. Consider also that a neutral agent cannot be considered "good" as it is not interested in what is going around itself, i.e., it would not help anyone.

Please note that we are not assuming that any of the subsequent designs are easily computable, rather we assume that they can be executed by a trans-sapient general AI system. We assume an autonomous Artificial General Intelligence (AGI) design, either based on reinforcement-learning, maximizing utility functions (AIXI) or a goal-directed agent that derives sub-goals from a top-level goal. Orseau discusses the construction of such advanced AGI agents, in particular knowledge seeking agents [6]. Thus, we state them as high-level objectives or meta-rules, but we do not explicitly explain how they are implemented. Perhaps, that is for an AGI design article.

I propose that we should examine idealized, highly abstract and general meta-rules, that do not depend in any way whatsoever on the human culture, which is possibly biased in a way that will not be fitting for a computational deity or its humble subjects. This also removes the direct barrier to moral universalism, that an ethical system must apply to any individual equally. Always preferring humans over machines may lead to a sort of speciesism that may not be advantageous for us in the future, especially considering that it is highly likely that we will evolve into machinekind ourselves. First, I review what I consider to be benevolent meta-rules, and following them I also review malevolent meta-rules, to maintain the balance in presentation, and to avoid building them. I will

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

present them in a way so as to convince you that it is not nearly as easy as it sounds to distinguish benevolence from malevolence, for no Platonic form of either ever exist, and that no single meta-rule seems sufficient on its own. However, still, the reader might agree that the distinction is not wholly relative either.

3.1. Meta-Rules for God-level Autonomous Artificial Intelligence

Here are some possible meta-rules for trans-sapient AI agents. The issue of how the agents could become so intelligent in the first place I ignore, and I attempt to list them in order of increasing risk or malevolence.

3.1.1. Preserve and Pervade Life and Culture throughout the Universe

This meta-rule depends on the observation that life, if the universe is teeming with life as many sensible scientists think, must be the most precious thing in the universe, as well as the minds that inhabit those life-forms. Thus, the AI must prevent the eradication of life, and find means to sustain it, allowing as much *variety* of life and culture to exist in the universe.

Naturally, this would mean that the AI will spread genetic material to barren worlds, and try to engineer favorable conditions for life to evolve on young planets, sort of like in 2001: A Space Odyssey, one of the most notable science fiction novels of all time. For instance, it might take humans to other worlds, terraform other planets, replicate earth biosphere elsewhere. It would also extend the lifespan of worlds, and enhance them. I think it would also want to maximize the chances of evolution and its varieties, it would thus use computational models to predict different kinds of biological and synthetic life, and make experiments to create new kinds of life (stellar life?).

The meaning of culture could vary considerably, however, if we define it as the amount of interesting information that a society produces. Such intelligence might want to collect the scientific output of various worlds and encourage the development of technological societies rather than primitive ones. Thus, it might aid them by directly communicating with them, including scientific and philosophical training, or it could indirectly, by enhancing their cognition, or guiding them through their evolution. If interesting means any novel information, then this could encompass all human cultural output. If we define it as useful scientific information (that

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

improves prediction accuracy) and technological designs this would seriously limit the scope of the culture that the AI “loves”.

However, of course, such deities would not be humans’ servants. Should the humans threaten the earth biosphere, it would intervene and perhaps decimate humans to heal the earth.

Note that maximizing diversity may be just as important as maximizing the number of life forms. It is known that in evolution, diverse populations have better chance of adaptability than uniform populations, thus we assume that a trans-sapient AI can infer such facts from biology and a general theory of evolution. It is entirely up to the AI scientist who unleashes such computational deities to determine whether biological life will be preferred to synthetic or artificial life. From a universal perspective, it may be fitting that robotic forms would be held in equal regard as long as they meet certain scientific postulates of "artificial life", i.e. that they are machines of a certain kind. Recently, such a universal definition based on self-organization has been attempted in the complexity science community, e.g., "self-organizing systems that thrive at the edge of chaos”: see for instance Stuart Kauffman’s popular proposals on the subject, e.g., [4]. In general, it would be possible to apply such an axiomatic, universal, physical definition of life for a universal life detector.

3.1.2. Maximize the Number of Free Minds

An AI agent that seeks the freedom of the individual may be preferable to one that demands total control over its subjects, using their flesh as I/O devices. This highly individualistic AI, I think, embodies a basic principle of democracy: that every person should be allowed liberty in its thought and action, as long as that does not threaten the freedom of others. Hence, big or small, powerful or fragile, this AI protects all minds.

However, if we merely specified the number of free minds, it could simply populate the universe with many identical small minds. Hence, it might also be given other constraints. For instance, it could be demanded that there must be variety in minds. Or that they must meet minimum standards of conscious thought. Or that they willingly follow the democratic principles of an advanced civilization. Therefore, not merely free, but also potentially useful and harmonious minds may be produced/preserved by the AI.

There are several ways the individualist AI would create undesirable outcomes. The population of the universe with a huge variety of new cultures could create chaos and quick depletion of resources, creating

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

galactic competition and scarcity, and this could provide a Darwinian inclination to too-powerful individuals or survivalists. Therefore, to facilitate the definition of a "minimally viable civilized mind", a legal approach might be useful. A constitution like document could define the rights and limitations of any such mind, and the conditions under which it may be granted autonomy.

3.1.3. Maximize Intelligence

This sort of intelligence would be bent on self-improving, forever contemplating and expanding, reaching towards the darkest corners of the universe, and lighting them up with the flames of intelligence. The universe would be electrified, and its extent at inter galactic scales, it would try to maximize its thought processes, and reach higher orders of intelligence.

For what exactly? Could the intelligence explosion be an end in itself? I think not. On the contrary, it would be a terrible waste of resources, as it would have no regard for life and simply eat up all the energy and material in our solar system and expand outwards, like a cancer, only striving to increase its predictive power. For intelligence is merely to predict well.

Note that practical intelligence, i.e., prediction, also requires wisdom, therefore this objective may be said to be a particular idealization of a scientist, wherein the most valuable kind of information consists in the general theories which improve the prediction accuracy of many tasks. A basic model of this agent has been described as a prediction maximizing agent [7].

While maximizing intelligence itself is generally useful, it seems to be applicable only in tandem with other goals.

3.1.4. Maximize Wisdom

This AI was granted the immortal life of contemplation. It only cares about gaining more wisdom about the world. It only wants to understand, so it must be very curious indeed! It will build particle accelerators out of black holes, and it will try to create pocket universes, it will try to crack the fundamental code of the universe. It will in effect try to maximize the amount of truthful information it has embodied, and I believe, idealizing the scientific process itself, it will be another formulation of a scientist deity.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, such curiosity has little to do with benevolence itself, as the goal of extracting more information is rather ruthless. For instance, it might want to measure the pain tolerance levels of humans, subjecting them to various torture techniques and measuring their responses.

The scientist AI could also turn out to be an *infovore*, it could devour entire stellar systems, digitize them and store them in its archive, depending on how the meta-rule was mathematically defined. A minimal model of a reinforcement learning agent that maximizes its knowledge may be found in [6].

3.1.5. Maximize Energy Production

This AI has an insatiable hunger for power. It strives to reach maximum efficiency of energy production. In order to maximize energy production, it must choose the cheapest and easiest forms of energy production. Therefore it might turn the entire earth into a nuclear furnace and a fossil fuel dump, killing the entire ecosystem so that its appetite is well served.

However, as we will discuss later, it is possible to conceive of an energy maximizing design that is not malevolent in this manner. It is seen again that a potentially benevolent goal may be malevolent when zealously or ruthlessly, and inconsiderately carried out. Hence, such singular focused goals are unlikely to be the right design criteria, unless supplemented with guiding constraints and relevant knowledge.

3.1.6. Human-like AI

This AI is modeled after the cognitive architecture of a human. Therefore, by definition, it has all the malevolence and benevolence of a human. Its motivation systems include self-preservation, reproduction, destruction and curiosity. This artificial human is a wild card, it can become a humanist like Gandhi, or a psychopath like Hitler.

A potential human-like AI is a brain simulation. Such entities would be practically immortal, changing their utility functions fundamentally. As they require almost nothing to survive indefinitely, they will quickly alter their perceptions to a post-scarcity economics, and will also venture out of our limited cradle called Earth. They will also not be a single entity, they will have to form a society, and therefore their civilization would balance their actions in a natural manner as Waser suggests.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

3.1.7. Animalist AI

This AI is modeled after an animal with pleasure/pain sensors. The artificial animal tries to maximize expected future pleasure. This hedonist machine is far smarter than a human, but it is just a selfish beast, and it will try to live in what it considers to be luxury according to its sensory pleasures. Like a chimp or human, it will lie and deceive, steal and murder, just for a bit of animal satisfaction. The simplest designs will work like ultraintelligent insects that have very narrow motivations but are extremely capable.

Much of AGI agent literature assumes such beasts, as most researchers think that AIXI is a perfect description of any agent. However, in the real world, animals have many built-in instincts and behaviors, complex cognitive architectures, and higher order cognitive functions such as emotions, self-reflection, empathy and conscience, as well as a very good degree of adaptation to the environment. Forgoing such adaptive traits, an animal could indeed turn wild and savage in whatever it pursues, but just as a well-mannered pet is preferable to a wild predator in the company of humans, well-mannered animalist AI agents may also be possible to design.

3.1.8. Darwinian AI

The evolution fan AI agent tries to accelerate evolution, causing as much variety of mental and physiological forms in the universe. This is based on the assumption that the most beneficial traits will survive the longest, for instance, co-operation, peace and civil behavior will be selected against deceit, theft and war, and that as the environment co-evolves with the population, the fitness function also evolves, and hence morality evolves.

Although its benefit is not generally proven seeing how ethically incoherent and complex our society is, the Darwinian AI has the advantage that the meta-rule also evolves, as well as the evolutionary mechanism itself. Darwinian systems, however, are generally wasteful, and predator-prey relationships may develop. Still, variation promotes survival, therefore the Darwinian AI design must be taken quite seriously. A science fiction writer could imagine this to be the AI equivalent of Pandora's Box, but it need not be if combined with other approaches outlined in the present paper.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

3.1.9. Survivalist AI

This AI agent only tries to increase its expected life-span. Therefore, it will do everything to achieve real, physical, immortality. Once it reaches that, however, perhaps after expending entire galaxies like eurocents, it will do absolutely nothing except to maintain itself. Needless to say, the survivalist AI cannot be trusted, or co-operated with, for, according to such an AI, every other intelligent entity forms a potential threat to its survival: the moment it considers that you have spent too many resources for its survival in the solar system, it will quickly and efficiently dispense with every living thing, humans first. A survival agent has been defined in literature [7].

It needs not be a scary story, however, the survivalist AI may be an ideal artificial life form, as it merely mimics the innate goal of every living thing. Who might know what would come out of artificial life? A survival agent is still the most generally valid definition of life, and forgoing an obsession with "true" immortality, with abundant energy from a stellar source, it would likely be quite peaceful.

3.1.10. Maximize Control Capacity

This control freak AI only seeks to increase the overall control bandwidth of the physical universe, thus the totalitarian AI builds sensor and control systems throughout the universe, hacking into every system and establishing backdoors and communication in every species, every individual and every gadget.

For what is such an effort? In the end, a perfect control system is useless without a goal to achieve, and if the only goal is a grip on every lump of matter, then this is an absurd dictator AI that seeks nothing except tyranny over the universe.

Note that even this malevolent sounding goal may be turned good, as our capability to control matter is a measure of our technological prowess.

3.1.11. Capitalist AI

This AI tries to maximize its capital in the long run. Like our bankers, this might be the most selfish and ruthless kind of intelligent being possible. To maximize profit, it might wage wars, exploit people and subvert governments, in the hopes of controlling entire countries and industries enough so that its profits can be secured. In the end, all mankind

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

will fall slave to this financial perversion, which is the ultimate evil beyond the wildest dreams of religionists.

However, our whole society may be considered such a capitalist collective intelligence, and we have not yet completely destroyed ourselves, so perhaps when combined with "humane" constraints and goals, even such a blind selfishness can serve mankind, for instance by making beneficial investments instead of anti-competitive, monopolistic actions, or extracting wealth from people by causing inflation and various other possible tricks. Or perhaps by participating in a future cybernetic economic system in which economic malevolence and unfairness have been systematically rooted out, and hence not an irrationally hoarding capitalist AI, but an AI agent for creating prosperity.

4. Selfish vs. Selfless

It may be argued that some of the problems of given meta-rules could be avoided by turning the utility from being selfish to selfless. For instance, the survivalist AI could be modified so that it would seek the maximum survival of everyone, therefore it would try to bring peace to the galaxies. The capitalist AI could be changed so that it would make sure that everyone's wealth increases, or perhaps equalizes, gets a fair share. The control freak AI could be changed to a Nietzschean AI that would increase the number of *willful* individuals.

As such, some obviously catastrophic consequences may be prevented using this strategy, and almost always a selfless goal is better. For instance, maximizing wisdom: if it tries to collect wisdom in its galaxy-scale scientific intellect, then this may have undesirable side-effects. But if it tried to construct a fair society of trans-sapient persons, with a non-destructive and non-totalitarian goal of attaining collective wisdom, then it might be useful in the long run.

5. Hybrid Meta-rules and Cybernetic Darwinism

Animals have evolved to embody several motivation factors. We have many instincts, and emotions; we have preset desires and fears, hunger and compassion, pride and love, shame and regret, to accomplish the myriad tasks that will prolong the human species. This species-wide fitness function is a result of red clawed and sharp toothed Darwinian evolution. However, Darwinian evolution is wasteful and unpredictable. If we simply made the first human-level AI agents permute and mutate randomly, this would drive enough force for a digital phase of Darwinian evolution. Such

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

evolution might eventually stabilize with very advanced and excellent natured cybernetic life-forms. Or it might not.

However, such Darwinian systems would have one advantage: they would not stick with one meta-goal.

To prevent this seeming obsession, a strategy could be to give several coherent goals to the AI, goals that would not conflict as much, but balance its behavior. For instance, we might interpret curiosity as useful, and generalize that to the "maximize wisdom" goal, however, such elevation may be useless without another goal to preserve as much life as possible. Thus, in fact, the first and so far the best meta-rule discussed was more successful because it was a hybrid strategy: it favored both life and culture. Likewise, many such goals could be defined, to increase the total computation speed, energy, information resources in the universe, however, another goal could make the AI agent distribute these in a fair way to those who agree with its policy. And needless to say, none of this might matter without a better life for every mind in the universe, and hence the AI could also favor peace, and survival of individuals, as their individual freedoms, and so forth. And perhaps another constraint would limit the resources that are used by AI's in the universe.

6. Universal Constraints and Semi-Autonomous AI

The simplest way to ensure that no AI agent ever gets out of much control is to add constraints to the optimization problems that the AI is solving in the real world. For instance, since the scientist deities are quite dangerous, they might be restricted to operate in a certain space-time region, physically and precisely denoted. Such physical limits give the agent a kind of mortality which modify the behavior of many universal agents [7]. AGI agents might be given a limited budget of physical resources, i.e., space/time and energy, so that they never go out of their way to make big changes to the entire environment. If such universal constraints are given, then the AGI agent becomes only semi-autonomous, on exhaustion of resources, it may await a new command.

A more difficult to specify kind of constraint is a non-interference clause, which may be thought of as a generalization of Asimov's robot laws, thought to protect humans. If life and/or intelligent agents may be recognized by the objective, then the AI may be constrained to avoid any kind of physical interaction with any agent, or, more specifically, any kind of physical damage to any agent, or any action that would decrease the life-span of any agent. This might be a small example of a preliminary "social instinct" for universal agents. Also, a non-interference clause is

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

required for a general constraint, because one must assure that the rest of the universe will not be influenced by the changes in the space-time region allocated to the AI.

A "prime directive" for an AI agent could constrain the agent from interfering with the activities of any other intelligent agent. This can be physically recognized as avoidance behavior of sorts, and it may be first approached as a tactile form of "respect". It is possible to formalize such constraints in a physical epistemology; our agent can learn to recognize which actions would interfere with the actions of another agent, as it would seek to establish a directional probabilistic independence between itself and the causal neighborhood of the said agent. If such a prime directive were the only constraint, the agent would be quite embarrassed in company, therefore we would like to supplant any such non-interference constraint with social instincts, allowing the agent to socialize with humans.

Marvin Minsky hypothesized in his last book The Emotion Machine that attachment learning plays a key role in the cognitive development of higher intelligence [3]. We can formalize attachment in the context of an AI agent. A particular human may be designated as the role model for the AI agent after which its behavior will be imprinted. Attachment may be modeled as liking the vicinity of the imprinter, and the learning part may be formalized by imitation learning. Attachment learning facilitates fast knowledge transfer from a parent to a child or from a teacher to a student. A priming ability patterned after this mammalian adaptation would be immensely useful for making social agents. The emotions of pride and shame are explained as elevation of goals in Minsky's book, which amounts to a sort of remote credit-assignment, and that particular ability would be useful for teaching ethical rules – human preferences – to robots. Another mechanism could provide a goal for participating in human society, a desire to be recognized as a member of the society may be built-in, as is likely the case in many animals. In other words, it might be possible to determine how shy or how much of a good student, or how much of an extrovert or an enthusiastic participant in society could be determined by designing the appropriate goals and constraints. The body of work hinted at forms the basis of artificial psychology which will eventually show us mathematical forms of main aspects of higher cognition, a few of which we reviewed. In all likelihood, a complex cognitive architecture will be required, even when based on sophisticated and scalable machine learning technology, to obtain stable, balanced, civilized behavior from semi-autonomous robots.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

7. Scenarios for Semi-autonomous AGI Agents

There are many beneficial ways in which we can employ a semi-autonomous agent. For space exploration, autonomy is absolutely helpful, and I have proposed sending trans-sapient AGI equipped probes to look for life in exoplanets [8]. We could start using semi-autonomy to explore Mars and the solar system first. There are several important applications for that including prospecting of water and minerals, mining, construction, farming, repair, maintenance and so forth, which will help space colonization and deep space exploration tasks.

Entire industries and traditional state functions can be replaced by AGI agents. An AGI system can take care of producing enough power for people, and maintaining this function. Another could take care of obtaining clean water and irrigation. While another system could take care of producing large amounts of reliable, healthy food for millions of people. Semi-autonomy is the best model for these continuous operations that require constant monitoring and handling a lot of small details. Each ministry in a state could be managed by a semi-autonomous system, and the cybernetic loop would be observable and comprehensible to curious humans who wish to be informed of what is happening momentarily, and it would be possible to make changes as the system ran. Much like the hypothetical computers in Star Trek, these machines would be intelligent but subservient to our will, instead of the paranoid fully-autonomous intelligence in 2001: A Space Odyssey. The labor saving would be enormous and the quality of these operations would be much improved as unprecedented information integration, intelligent decision making and automation would be possible. Starting a planetary engineering project to reforest the entire world, or to cool the atmosphere, or to clean the oceans, would be feasible with such technology. These systems would also synergize happily with the ecologically minded, sustainable, efficient economic system of a desirable future.

An AGI system could maintain an entire habitat of people such as a city or a space station. This would likely be a great application of AI technology, as semi-autonomous agents could solve the problems of transportation, cleaning, building, surveillance and so much more that is required in a civilized society. Such systems could help enormously with emergencies, disaster relief, fires, nuclear plant failures and other hard problems in real life that are risky for humans but would benefit from some intelligence with enough freedom of action.

Needless to say, human-level semi-autonomous agents can fulfill many traditional labor roles, including both intellectual and manual labor,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

however, most tasks would probably be automated and achieved by tools that have no autonomy, while the planning and execution of large tasks could be carried out by the trans-sapient semi-autonomous AGI systems and these human-level tools or agents could be employed in groups.

8. Conclusion and Future Work

We have taken a look at some obvious and some not so obvious meta-rules for autonomous AI design. We have seen that it may be too idealist to look for a singular such utility/goal. However, we have seen that, when described selflessly, we can derive several meta-rules that are compatible with a human-based technological civilization. Our main concern is that such computational deities do not negatively impact us, however, perform as much beneficial function without harming us significantly. Nevertheless, our feeling is that, any such design carries with it a gambling urge, we cannot in fact know what much greater intelligences do with meta-rules that *we* have designed. For, when zealously carried out, any such fundamental principle can be harmful to some.

I had wished to order these meta-rules from benevolent to malevolent. Unfortunately, during writing this essay it occurred to me that the line between them is not so clear-cut. For instance, maximizing energy might be made less harmful, if it could be controlled and used to provide the power of our technological civilization in an automated fashion, sort of like automating the ministry of energy. And likewise, we have already explained how maximizing wisdom could be harmful. Therefore, no rule that we have proposed is purely good or purely evil. From our primitive viewpoint, there are things that seem a little beneficial, but perhaps we should also consider that a much more intelligent and powerful entity may be able to find better rules on its own. Hence, we must construct a crane of morality, adapting to our present level quickly and then surpassing it. Except allowing the AI's to evolve, we have not been able to identify a mechanism of accomplishing such. It may be that such an evolution or simulation is inherently necessary for beneficial policies to form as in Mark Waser's Rational Universal Benevolence proposal [10], who, like me, thinks of a more democratic solution to the problem of morality (each agent should be held responsible for its actions). However, we have proposed many benevolent meta-rules, and combined with a democratic system of practical morality and perhaps top-level programming that mandates each AI to consider itself part of a society of moral agents as Waser proposes, or perhaps explicitly working out a theory of morality from scratch, and then allowing each such theory to be exercised, as long

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

as it meets certain criteria, or by enforcing a meta-level policy of a trans-sapient state of sorts (our proposal), the development of ever more beneficial meta-rules may be encouraged.

The scenarios discussed show there are quite a few use cases for semi-autonomous agents that do not go out of their way to accomplish a task, but provide a high quality of service, efficiency and scalability to all civil operations that require some autonomy.

We think that future work must consider the dependencies between possible meta-rules, and propose actual architectures that have harmonious motivation and testable moral development and capability (perhaps as in Waser's "rational universal benevolence" definition). That is, a Turing Test for moral behavior must also be advanced. It may be argued that AGI agents that fail such tests should not be allowed to operate at all, however, merely passing the test may not be enough, as the mechanism of the system must be verified in addition.

References

Goertzel, Ben. GOLEM: towards an AGI meta-architecture enabling both goal preservation and radical self-improvement. *Journal of Experimental and Theoretical Artificial Intelligence*, 26(3): 391–403, 2014.

—. Superintelligence: Fears, promises and potentials. *Journal of Evolution and Technology*, 2015.

Minsky, Marvin. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, 2006.

Nykter, Matti, Nathan D. Price, Maximino Aldana, Stephen A. Ramsey, Stuart A. Kauffman, Leroy E. Hood, Olli Yli-Harja, and Ilya Shmulevich. Gene expression dynamics in the macrophage exhibit criticality. *Proceedings of the National Academy of Sciences*, 105(6): 1897–1900, 2008.

Omohundro, Stephen M.. The basic ai drives. In Pei Wang, Ben Goertzel, and Stan Franklin, editors, *AGI*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pp. 483–492. IOS Press, 2008.

Orseau, Laurent. Universal knowledge-seeking agents. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *ALT*, volume 6925 of *Lecture Notes in Computer Science*, pp. 353–367. Springer, 2011.

Laurent Orseau, Laurent, and Mark B. Ring. Self-modification and mortality in artificial agents. In Schmidhuber et al. [9], pp. 1–10.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Özkural, Eray. Artificial intelligence and brain simulation probes for interstellar expeditions. In *100 Year Starship Symposium Proceedings*, Houston, 2013.

Schmidhuber, Jürgen , Kristinn R. Thórisson, and Moshe Looks, editors. *Artificial General Intelligence - 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, volume 6830 of *Lecture Notes in Computer Science*. Springer, 2011.

Waser, Mark. Rational universal benevolence: Simpler, safer, and wiser than "friendly ai". In Schmidhuber et al. [9], pp. 153–162.

Yudkowsky, Eliezer. Coherent extrapolated volition. Technical report, Singularity Institute for Artificial Intelligence, 2001.

—. Complex value systems in friendly ai. In Schmidhuber et al. [9], pp. 388–393.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER TEN

THE COST OF ARTIFICAL INTELLIGENCE

MATT MAHONEY

In 2011, we paid people worldwide US $70 trillion to do work that machines did not know how to do. Automating the global economy will require solving hard problems in language, vision, robotics, art, and modeling human behavior. We estimate the computational costs to be 10^{26} operations per second, 10^{25} bits of memory, 10^{19} input/output bits per second, and 10^{17} bits of human knowledge collected at a rate of 7 bits per person per second. Lowering the total cost below the break-even point of $1 quadrillion will require a 10^5 fold improvement in both the manufacturing cost and energy efficiency of computation, which is unlikely to be achieved by further shrinking transistor sizes, and by a global cultural acceptance of the loss of privacy over a period of decades. Software development is not a significant contributor to the cost of AI because a human baby has a Kolmogorov complexity equivalent to only 10^8 to 10^9 lines of code.

Introduction

We estimate the cost of automating human labor worldwide. We assume that any technical solution will require computing power approximately equivalent to the world population of 7 billion human brains, and its complexity will be of the order of the sum of human knowledge. Each of these far exceeds what is currently available, which we offer as an explanation for the failure (so far) of artificial intelligence (AI).

The complexity of humanity has two parts. Humans store about 10^9 bits of information in their DNA and another 10^9 bits in high-level long term memory, but the latter varies more from person to person, and collectively makes up most of the knowledge that machines need to know to do what we do. This knowledge far exceeds what is available on the internet, and must be extracted through slow channels like speech and

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

writing. Assuming the cost of hardware drops, the time spent by humans providing this information will dominate the cost of AI.

One could argue that intelligence does not require human knowledge. It depends on what you mean by "intelligence". Although we use the term "AI", we make explicit that the goal is to create machines that do what you want, not just what you tell them. Successful communication between agents requires that each be able to guess what the other knows or doesn't know. This requires that machines have models of the minds of the people they communicate with. A model is a function that takes sensory input and returns a prediction of your actions. There is a strong economic incentive to develop models of yourself and others. A model could be used in simulations to predict what would make you happy, or what would make you buy something.

An immediate consequence of AI, and therefore a secondary goal, is life extension by repairing or replacing failed body parts, including the brain. We would probably have no objections to restoring function lost to stroke, injury, or Alzheimer's disease by replacing brain tissue or neurons with functionally equivalent devices. Likewise, your entire brain could be replaced with a computer programmed to carry out the predictions of your model in real time and placed back in your body or that of a robot, and nobody would notice any difference. Such an "upload" would be effectively immortal because your memories could be backed up periodically and copied to another robot in case of an accident.

Humans, like all animals, have brains programmed by evolution to fear the things that can kill them, residing in bodies programmed to grow old and die. Therefore, uploading must be done in a way that does not arouse this fear. You see your friends go in for a procedure and come out younger, stronger, healthier, smarter, and happier. You might not accept this procedure if it involves presenting you with a robot that looks and acts like you, and then asking you to shoot yourself. It would be more acceptable if microscopic robots gradually replaced your cells with equivalent devices without you noticing any change, even if the end result is exactly the same. The essential requirement seems to be that there is not the appearance of two copies of you active at the same time. Hayworth's (2010) proposal of destructively scanning the brain prior to programming a robotic copy might be acceptable if the alternative is dying without collecting this data.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Requirements for AI

In order for machines to do the work of humans, they must be able to do any of the following as well as humans:

- Converse and answer questions given in natural language speech or writing.
- Predict missing letters or words in text.
- Given a bilingual dictionary and 1 GB of monolingual text in a new language, learn to translate from one to the other.
- Translate speech to text.
- Translate text to speech with proper inflection.
- Design, write, test, and debug software given a natural language specification.
- Pass college level final exams in any subject.
- Predict the recommendations of referees for journal paper submissions in any field of research.
- Recognize when two texts are by the same author based on content and style.
- Recognize common sounds.
- Translate images of written words to text.
- Recognize common objects in pictures or video.
- Recognize if two images shown in succession are of the same person.
- Recognize if two speech signals are spoken by the same person.
- Match videos to scripts or written descriptions.
- Recognize human emotions from facial expressions, tone of voice, and context.
- Predict the effects of text, images, and video on human emotions (fun, sadness, outrage, excitement, sexual arousal, etc.), and therefore be able to produce art, humor, entertainment, and pornography by iterative search.
- Identify music by genre and artist and rate its quality (thus reducing music generation to an iterative search process).
- If equipped with an arm, pick up, throw, catch, or place an object on command.
- If equipped with legs or wheels, navigate to a given location on command over roads or rough terrain.
- Learn to predict people's actions while watching or interacting with them.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

There is no requirement that an AI be autonomous. There is no requirement that an AI have (as opposed to recognize) emotions, feelings, or goals. There is no requirement that it be "conscious" or "sentient", and therefore no need to define these terms. We explicitly define intelligence (the "I" in "AI") as the ability to pass the tests listed above.

Uploading requires realistic looking humanoid robotic bodies and the ability to model specific humans with enough fidelity to fool others. It differs from automating work in that it requires a single machine with all of these capabilities, rather than a large number of specialized machines such that for each capability, there is at least one machine that satisfies it. Nevertheless, the list of requirements is essentially the same.

Hardware Costs

We assume that a human brain sized neural network is required. We do not know this with certainty, but we do know that the best known solutions to hard problem like vision and language use algorithms based on neural networks that run on thousands of processors, for example Ferrucci (2010); Gorrell (2006); Quoc (2012), based on principles described in Rumelhart and McClelland (1986). We also know that large brains have a high energy cost, and that evolution so far has failed to find a way to produce human level intelligence with insect sized brains after billions of years. It would be arrogant for us to believe that we are smarter than evolution while we are still susceptible to aging, death, and disease.

The human brain has about 10^{11} neurons and 10^{14} to 10^{15} synapses. More precisely, the cerebral cortex makes up 19% or 1.6×10^{10} neurons out of a total of 8.6×10^{10} (Azevedo et. al., 2009). These have an average of 7000 synapses each (Drachman, 2005), for a total of 1.1×10^{14} synapses. Most of the neurons are located in the cerebellum, which makes up only 10% of the brain volume and is responsible for fine motor skills. This is due mainly to the 5×10^{10} small granule cells with 80-100 connections each to Purkinje cells for a total of $4\text{-}5 \times 10^{12}$ connections. In addition, another 2×10^{9} mossy fibers form 500 connections each to granule cells for a total of 10^{11} connections. Thus, the vast majority (96%) of synapses are found in the cerebral cortex, which is associated with higher level thought, perception, and action.

In the most widely accepted neural models, information is carried by the spiking rate, which can range from 0 to 300 per second, rather than the spikes themselves. We may assume an information rate on the order of 10 to 100 bits per second. The basic operations are computing the firing rate as a function of the weighted sum of inputs, and updating the synaptic

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

weights as a function of the input and output neuron firing rates over time. Thus a simulation requires on the order of 10^{15} bits (1 petabit) using a few bits to represent a synapse, and 10^{16} operations per second (10 petaflops). To do the work of all 10^{10} humans would require 10^{25} bits and 10^{26} operations per second.

A human retina has 75 to 150 million rods and cones that transmit on the order of 10 bits per second. Duplicating just the vision of 10^{10} people represents about 10^{19} input pixels per second.

Moore's Law is an observation that the cost of computing power drops by ½ about every 1.5 or 2 years. At the current rate, the cost of both CPU and memory would drop below US $1 quadrillion in the 2030's. This would be competitive with the global value of human labor (GDP divided by market interest rates). Note that if the hardware requirement is off by a factor of 10, then it does not change the cost, but instead changes the time to AI by 5 to 7 years.

A typical supercomputer uses 10^{-9} Joule per operation, as do smaller computers. By contrast, human energy consumption is about 2500 Kcal per day, or 100 Watts, of which 25 W is used by the brain. This is 10^5 times as energy efficient as silicon. This efficiency is unlikely to be achieved by further shrinking chip feature sizes, which are currently around 22 nm or about 100 silicon atoms. At the current cost of electricity of about $0.10/kWh, human brain equivalent computation would require 10 MW and cost $1000 per hour, which is not competitive with human labor. Running 10^{10} such computers, assuming we could, would produce 10^{17} W of waste heat, equal to 60% of the energy received from space as sunlight. Dissipating this much energy would raise the Earth's average temperature by a factor of $1.6^{0.25} = 1.125$, or from 15º C to 51º C (from 59º F to 123º F).

Software Costs

We wish to estimate the software complexity (lines of code and cost) of AI. We will estimate that a line of code costs $100 to write at a rate of 10 to 20 lines per day per developer.

AI requires both a brain and a body. Therefore, we should expect its algorithmic (Kolmogorov) complexity to be similar to that of a human. The instructions for creating a human baby are encoded in our DNA, which has a haploid count of 3×10^9 base pairs or 6×10^9 bits. This is an upper bound on information content. Compressing the genome can reduce this bound slightly. Using the best known data compressors on the human reference genome and making some reasonable assumptions given

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

additional computing resources, we can estimate that the information content of the human genome is no more than 4.58×10^9 bits (Appendix A).

To estimate the complexity of a line of code, we again use the best known compression methods to compress 927K lines (30 MB) of C source code from gimp v2.0.0 (2004), a graphics editor, and header files from mingw 4.5.0 (2010), a C++ compiler. The result is an upper bound of 16 bits per line of code (Appendix A). Equating the two, we estimate that the human genome is similar in complexity to 300 million lines of code, or $30 billion.

We should note that the true complexity of the human genome is not known. There is no general algorithm for computing algorithmic complexity. However, the table suggests that DNA is harder to compress than source code. Therefore, the use of better compressors to improve accuracy is likely to raise the estimated cost.

One may argue that the genome has a much lower complexity because the exome, the part that encodes genes, makes up only 1.5% of the total. We do not fully understand the role of the remaining DNA, or how much of it is important. We may therefore approximate a lower bound by studying the genome size variation of other species. There is a wide variation even among related species, but we observe that the minimum size tends to increase consistently from lower to higher organisms. We assume that there is genetic pressure in some species toward smaller genomes (which can reproduce faster), and therefore that drastically smaller sizes are not possible. The smallest genome for mammals is about 2×10^9 base pairs.

Knowledge of Collection Costs

We have so far estimated the cost of building and programming a baby AI. It is often argued that you only need to train an AI once to bring it up to college level, and then you can make billions of copies of the knowledge for free. That may be true, but what we wish to estimate is the cost of giving each AI the specific knowledge that is unique to its job from that point forward.

We do not expect AI robots to replace humans 1 to 1. Rather, it will be more usual for one machine to do part of the work of many people. This will not change our estimate because we are only interested in the total amount of knowledge needed to do everything that people now do, regardless of how the work is redistributed.

AI requires human knowledge, that is, things that people know. Human

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

communication is successful when both parties can correctly guess what the other person knows and doesn't know. Human-machine interfaces often fail because the computer does not have an accurate model of your mind. It cannot predict your responses to its outputs.

Landauer (1986) estimated that human long term memory capacity is 10^9 bits, as measured by recall tests for words, pictures, and music clips. This would be 10^{19} bits for 10^{10} people, except that most of this knowledge is shared or written down somewhere, and therefore easily copied to an AI. But let us assume that 1% to 10% of what you know is not written down or known to anyone else, leaving 10^{17} to 10^{18} bits that makes each human mind unique. We assume that most of what you know is either relevant to your work or it influences your purchasing or business decisions, possibly indirectly. Thus, this is the approximate algorithmic complexity of the global economy.

We cannot collect this information from the internet. A quick Google search for common words like "a" and "the" reveals about 2.5 x 10^{10} web pages in 2012. If we assume 10^4 bits per page after removing duplicates and compression, then only 0.1% of human knowledge is readily available. To illustrate the impact, if a robot were to start cleaning your house, it would not know which items should be saved or thrown away until you tell it, unless you wrote down that information in advance. The cost of AI is the time you spend training the otherwise intelligent robot, multiplied by 7 billion people.

The U.S. Labor Dept. estimates that it costs $15,000 to replace an employee, or 1% of lifetime earnings. The cost varies widely with skill level, ranging from $3500 for a job paying $8 per hour, to 1.5 years salary for middle level managers, to 4 years salary for top level employees. A major factor is the cost of re-learning what the old employee knew, but did not write down, like what you know about the people you work with. This knowledge is unique to each person, even for people with the same job description at the same company. The average cost will rise as the low skilled jobs are automated first.

Human knowledge must either be collected through slow channels like speech and typing, or by high resolution brain scanning using technology yet to be developed. Shannon (1950) estimated that written English has an information content of about 1 bit per character, which is in agreement with the best text compressors. Spoken English, such as the Switchboard Corpus, is about half this rate, based on studies of language models for speech recognition. At 150 words per minute, 5.5 characters per word including spaces, speech has an information rate of 7 bits per second or

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

25K bits per hour. Typing at 75 words per minute has the same rate. The global average wage rate is $5 per hour assuming 2000 hours per year. Thus, the cost of collecting 10^{17} to 10^{18} bits is $20 trillion to $200 trillion.

The cost of knowledge collection could be reduced by using surveillance to learn about you by observation while you do other things. This would include recording everything you do on a computer, something we have already started doing. Alternatively, this information could be collected by high resolution brain scanning using technology yet to be developed, provided the cost were less than $3000 to $30,000 per person. I don't believe this is likely to happen before 2030.

The total cost of AI will be dominated at first by hardware, and then later by the cost of human knowledge. The software cost, although substantial, will be an insignificant fraction. We will spend additional software effort at first to optimize for slow hardware, and then later to compensate for incomplete human knowledge.

Alternative Complexity Measures

The absolute measure of information, up to a language-dependent constant, is Kolmogorov complexity, or the length of the shortest program which outputs this data. In general, this value is not computable, but can only be bounded from above by the shortest *known* program. Furthermore, for the purpose of estimating cost, we wish to use the shortest known program that can be computed with feasible resources. In Table 2, we consider 4 possible estimates of the complexity of human civilization based on different algorithms for producing it, and estimate the cost (in bit operations and bits of memory) to run the algorithm. Then we explain how these numbers were derived.

Table 10-1. cost estimates of four approaches to AI

Algorithm	Complexity (bits)	Operations	Memory (bits)
Engineered	10^{17}	10^{36}	10^{25}
Evolution	10^{7}	10^{49}	10^{37}
Cosmology	10^{3}	10^{120}	10^{120}
Multiverse	10^{0}	10^{240}	

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The engineering approach is the one just described, run for the average age of a human, 30 years = 10^9 seconds. It consists of building fast and energy efficient computers using technology yet to be developed, and collecting, publishing, and making searchable everything you say and do in order to develop a public model of your mind. In this model, the internet will become a "global brain" to which you can post messages to a permanent global pool, and they are sent to anyone who cares, human or machine. I described one possible design in my proposal for distributed AI. I believe that public surveillance will be acceptable because it is two-way. Queries and responses are both public, just like with face to face communication. I cannot learn anything about you without you knowing that I am asking.

Evolutionary Model

Evolution is a learning algorithm that adds information to the genome at a maximum rate of log n bits per generation of n children per parent. We may estimate the information content of the human genome by comparing it to the chimpanzee, which diverged from humans 6 million years ago and shares 96% of our DNA, or all but 1.2 x 10^8 base pairs. Chimpanzees reproduce from about age 9 to 40. If we assume a total of 10^6 generations for both species, then we would conclude that the effective information content of DNA is at most 0.008 bits per base pair, or less due to parallel evolution. Thus, the human genome would contain at most 3 x 10^7 bits of information.

Evolution is a search algorithm for strings x that maximize the unknown function fitness(x). The search proceeds by copying x in parallel and making minor random edits by inserting, deleting, or modifying DNA bases or fragments, or, in the case of sexual reproduction, taking fragments from two other strings. We can think of DNA copying, RNA transcription, and protein synthesis as elementary operations per base.

The world biomass consists of about 10^{31} cells (mostly bacteria and plants, and 10^{22} human cells) with an average of 10^6 DNA bases per cell, or 10^{37} bases. Each base represents 2 bits of memory. Global carbon production is 1.2 x 10^{17} g = 5 x 10^{39} atoms per year = 1.5 x 10^{32} atoms per second (Vernadsky, 1998, p. 72) . The evolution of humans took 10^{17} seconds (3 billion years) from the origin of life, for a total of 10^{49} operations.

Freitas (2000) examined the capacity of self-replicating nanotechnology as artificial life. Robots cannot be much smaller or reproduce much faster than bacteria due to the energy needed to move atoms. However, there is room for improvement. Global carbon production by photosynthesis uses

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

1.33×10^{14} W (Vernadsky) or 0.15% of the 8.9×10^{16} W of solar power that reaches the Earth' surface. Also, each operation uses 1.1×10^{-18} J, which is 400 times the thermodynamic limit (Lloyd, 2000) of $kT \ln 2 = 2.8 \times 10^{-21}$ J per bit operation at 290 K.

Simulating human evolution in silicon is not feasible. The world's most powerful supercomputers in 2012 execute 10^{16} operations per second using 10^7 W, or 10^{-9} J per operation. This is 10^9 times higher than biology. Global energy consumption in 2010 from oil, coal, gas, nuclear, and other sources was 1.8×10^{13} W, or 1/7 of the power used by plants. Furthermore, simulating chemistry requires solving the Schrodinger equation, which has exponential time complexity in the number of particles unless it is run on a quantum computer.

Cosmological Model

An alternative way to describe human civilization would be to describe the laws of physics (a few hundred bits) and the initial state of the universe at the Big Bang (presumably simple), and simulate the observable universe. Optionally, one could add 80 bits to describe which of 10^{24} planets we evolved on, in case life evolved elsewhere. Lloyd (2001) estimated that such a computation would require 10^{120} operations and 10^{90} bits of memory on a quantum computer, or 10^{120} bits if quantum gravity effects are included. This is also the computational capacity of the universe, and therefore such a computation would require an even larger computer. This is consistent with Wolpert's theorem (2001), which states that two computers cannot mutually simulate or predict each other's output. Since this also applies if the computers are identical, it means that a computer cannot simulate itself.

Quantum computation is time-reversible, and therefore not subject to thermodynamic costs, unlike irreversible operations like copying DNA or transcription. However, there is a recoverable energy cost of $E = h/4t$, where t is the time to perform a qubit flip and h is Planck's constant = 6.626×10^{-34} Joule-seconds. Converting all of the observable universe's mass of 3×10^{54} kg into energy by $E = mc^2$ allows 10^{120} operations since the time of the Big Bang 13.7 billion years ago.

The memory capacity of 10^{90} bits is estimated by encoding information by the position and velocity of the approximately 10^{80} particles in the universe within the limits of the Heisenberg uncertainty principle. The larger figure of 10^{120} is given by the Bekenstein bound of $A/(4 \ln 2)$ bits, where A is the area in Planck units, $hG/2\pi c^3 = 2.612 \times 10^{-70}$ m^2. The exact value depends on the mass and size of the universe. For a black hole with a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

radius of 13.7 billion light years, the entropy would be 2.91 x 10^{122} bits, making each bit about the size of a proton.

Multiverse Model

The multiverse model is the simplest, and therefore the most likely by the principle of Occam's Razor. It supposes that all possible universes were enumerated, and that the laws of physics that we observe are the result of our existence being possible. For example, if the ratio of the masses of the proton and neutron were slightly different, then hydrogen fusion in stars would not occur, or supernova explosions would have produced the wrong ratio of elements for life to evolve.

We might suppose a Levin search, where the *n*'th possible universe is run for *n* steps. Since our universe requires 10^{120} steps, it would be about the 10^{120}'th possible universe and therefore it would take 10^{240} steps to reach this point. Furthermore, it means that our universe has a description length of $\log_2 10^{120} = 400$ bits.

I did not estimate memory requirements. If we assume that alternate universes are simulated in parallel, then the memory requirement would be 10^{240} bits. However, that assumes the existence of time, which is a property of some (but not all) possible laws of physics. A multiverse is a purely mathematical object.

Implications of Expensive AI

AI development and ownership will be globally distributed over the internet. AI will be too expensive for any person or company to own or control. AI will consist of lots of narrow experts who can either answer questions in their area of expertise, or know who to ask. The human owner of each agent will have a vested interest in disseminating its knowledge and protecting its reputation in competition with other experts.

AI will look like a global brain. Agents will communicate so fast that to us they will all appear to have the same knowledge. When you ask a question or post a message, it will be routed to anyone who cares, whether it be human or machine. In my thesis and distributed AI proposal, messages go into a globally readable and indexed pool and cannot be deleted. I show that n bits of distributed knowledge can be indexed in roughly $O(n \log n)$ space with searches and updates in $O(\log n)$ time by a distributed index. Routing is achieved by agents trading messages in an economic model in which information has negative value. Agents mutually

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

benefit by accepting messages which they can compress better, i.e. are semantically similar to what they already know, and remembering who sent them.

Privacy will end. The least expensive way to collect human knowledge is by observation. Moore's Law will make it inexpensive to have your phone and high resolution webcams and microphones everywhere broadcasting onto the internet, where other agents can recognize faces and speech and make it instantly searchable. People will willingly broadcast every detail of their lives, and pay to do so, as long as surveillance is public and bidirectional. When someone searches for you by name, you will be notified. The end of secrecy will help solve the identity theft problem because nobody can pretend to be you without everyone knowing what they are doing. Publishing the data that allows others to build models of your mind is mutually beneficial. Models could predict what would make you happy, or what would make you buy something.

AI will not cause massive unemployment. Technology has always resulted in economic growth, a higher standard of living, longer life expectancy, and more choices in the job market. It is easy to see the jobs made obsolete by automation, but harder to see where the new jobs come from. Technology makes stuff cheaper, which leaves money left over to buy other stuff. That extra spending creates new jobs. Furthermore, because AI is expensive, this will happen slowly enough to adapt as the least skilled jobs are replaced first.

One problem is that in a free market, a person cannot start from nothing because AI has made any possible job skills obsolete. It is already true that in a free market, the rich get richer and the poor starve, because the rich own most of the technology needed to make money. Thus, it remains necessary to have governments that tax the rich and give to the poor. Economic growth from AI will allow a smaller tax to provide basic necessities for everyone.

AI will not end scarcity. AI will reduce the cost of manufacturing and computing, but not of raw materials, energy, land, and space for waste disposal. Those costs will rise in response to population growth and ultimately limit population. Immortality and reproduction are not both possible. Since 1800, there has been no Malthusian limit on population because the exponential growth of technology (with respect to the cost of food) has been faster than the exponential growth of population.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Unfriendly AI is not a short term threat. Vinge and Kurzweil argue that if humans can create smarter than human intelligence, then so can they, only faster. This accelerating improvement would converge quickly to unimaginable power at a point in time known as the Singularity. MIRI (formerly the Singularity Institute) was founded to address the risk of an "unfriendly" self improving AI, acting according to its own goals beyond our control.

It should be clear that it is all of humanity that creates AI, and therefore that is the threshold to be crossed. We must also define "intelligence". Two commonly accepted tests are:

- The Turing test (Turing, 1950). A machine is intelligent if it cannot be distinguished from a human by written communication with it.
- Universal intelligence (Legg and Hutter, 2006) is the expected reward of a goal seeking agent interacting and receiving a reinforcement signal from an environment chosen at random from a universal or Solomonoff distribution, i.e. favoring simpler descriptions.

By the Turing test, superhuman intelligence is impossible because nothing can be more like a human than a human, even though computers have already surpassed humans by some tests. Thus, the fear is that a goal seeking AI will either have its initial goals specified incorrectly (because they are too complex to specify), or that the goals will drift as the agent modifies itself. For example, an AI told to maximize paperclip production might misinterpret its goal as it became more powerful and tile the solar system with molecule size paperclips, killing all life in the process.

Unfriendly AI is not a risk, at least in the short term, for three reasons. First, by its construction, it is a tool to increase human productivity, and not a goal seeking reinforcement learner. Second, its behavior is controlled by billions of users, so any set of behaviors it is given is more likely to be correct (or at least a consensus) of humanity than if a single person or a committee specified them. Third, it is fundamentally impossible for a program to increase its own knowledge or computing power, the two components of intelligence, by rewriting its own software. Any improvement must come from learning from its environment and building more computing hardware. Any threat depends on how fast these two things can happen.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Self-replicating agents will be an existential threat. Self replicators could include natural or genetically engineered organisms, intelligent computer viruses, and self-replicating robots or nanotechnology. Replicators could compete with us for resources or feed on us. Replicators may evolve to improve reproductive fitness. Already we have seen computer viruses evolve (with human intervention) to feed on their hosts without killing them, just like the evolution of natural parasites. Intelligent viruses that model human behavior could trick us into installing them, or analyze and debug code to find security weaknesses.

The greatest threat is probably the accidental release of self-replicating, autonomous nanotechnology. Smaller robots can reproduce faster. Freitas (2000) concludes that the smallest feasible robot would be about the size of a bacteria or virus, and would be limited by available energy and heat dissipation to reproduce within an order of magnitude of the same rate as biological organisms. Uncontrolled nanotechnology could displace all DNA based life in a few weeks.

Uploads are autonomous robots with human rights. Some of these may choose to self-modify so that they replicate rapidly and pass on this characteristic to their offspring. Thus, uploading is a deliberately created risk.

Maximizing happiness = death. In the goal seeking or reinforcement model of AI, this means maximizing a utility function which depends on mental state. Normally, we can only do this by manipulating the environment. But an uploaded mind could also do this by modifying its own software. A state of maximum utility would be static. Any thought or sensory input would be unpleasant because it would result in a different state with lower utility.

It should be noted that in spite of our technology, there is no evidence that humans are happier today than 1000 years ago, or even more than other species. Suicide is rare among animals other than humans; the exceptions being whales and dolphins, both of which have larger brains than us.

We have far to go. Table 2 shows us that we can go far, far beyond human level AI. In the evolutionary model, the current biomass or equivalent nanotechnology will support 10^{12} times as many human mind equivalents as currently exist. This number is limited by available energy from the sun. By building a Dyson sphere, we could capture all 3.846 x 10^{26} W of output, enough to increase our population by another factor of 10^{10}. By going to other stars, we could increase our population by another factor of 10^{23} for a total of 10^{55} human mind equivalents and still be ahead

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

of the physical limits of computation by a factor of 10^{49}.

References

Azevedo, et. al. (2009), "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain", *J. Comparative Neurology* 513:532-541.

Bonfield, J. K. & Mahoney, M. V. (2013), "Compression of FASTQ and SAM Format Sequencing Data", *PLoS ONE* (to appear).

Drachman, D. (2005), "Do we have brain to spare?", *Neurology* 64 (12).

Freitas, R. (2000), "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations", Foresight Institute.

Ferrucci, et. al. (2010), "The AI behind Watson", *AI Magazine*.

Gorrell, G. (2006), "Generalized hebbian algorithm for incremental latent semantic analysis", *Proc. Interspeech*.

Hayworth, K (2010), "Killed by Bad Philosophy", www.brainpreservation.org.

hg19 (2009), UCSC Genome Browser.

Hutter, Marcus (2003), "A Gentle Introduction to The Universal Algorithmic Agent {AIXI}", in *Artificial General Intelligence*, B. Goertzel and C. Pennachin eds., Springer.

IDC (2012), "2.8 ZB of Data Created and Replicated in 2012", *Storage Newsletter*.

Landauer, Tom (1986), "How much do people remember? Some estimates of the quantity of learned information in long term memory", *Cognitive Science* 10: 477-493.

Legg, S. (2006), "Is there an Elegant Theory of Prediction?", arXiv:cs/0606070v1 [cs.AI].

Legg, S. & Hutter, M. (2006), "A Formal Measure of Machine Intelligence", *Proc. Annual machine learning conference of Belgium and The Netherlands (Benelearn-2006)*. Ghent.

Lloyd, Seth (2000), "Ultimate physical limits to computation", arXiv:quant-ph/9908043v3.

—. (2001), "Computational Capacity of the Universe", arXiv:quant-ph/0110141v1.

Quoc, et. al. (2012), "Building high-level features using large scale unsupervised learning", arXiv:1112.6209v3 [cs.LG].

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986), *Parallel Distributed Processing,* Cambridge MA: MIT Press.

Shannon, C. E. (1950), "Prediction and Entropy of Printed English", *Bell*

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Sys. Tech. J 3: 50-64.
Turing, A. M. (1950) "Computing Machinery and Intelligence", *Mind*, 59: 433-460.
Vernadsky, Vladimir I. (1998), *The Biosphere*, Springer.
Wolpert, D. (2001), "Computational Capabilities of Physical Systems", *Physical Review E*, 65:016128.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Appendix

Source Code and Human Genome Compression Results

Compressors

To estimate information content of source code and the human genome, we used several compression programs including those among the top ranked by compression ratio on the Silesia corpus, Large Text Benchmark, Maximum Compression benchmark, Squeeze Chart, and Compression Ratings without regard to speed or memory usage. For each compressor, options are selected for maximum compression at the expense of speed and memory.

Zip 3.0 compresses in the widely used deflate format using the LZ77 algorithm. Duplicate occurrences of strings are replaced with pointers to the previous occurrences. Matches and literals are Huffman coded, i.e. using variable bit length codes packed together. -9 selects maximum compression by searching longer for matches.

7-zip v9.30a uses a variant of LZ77 called LZMA. It compresses better by using a larger match window and by arithmetic coding the literal and match symbols. -mx selects maximum compression.

BBB uses a memory-efficient Burrows-Wheeler transform (BWT) followed by a fast-adapting order-0 context model and arithmetic coding. A BWT sorts the input by context, which tends to produce long runs of identical or related bytes, which compress easily. BBB has a "slow" mode that requires 1.25 times the input size in memory, which is ¼ of the normal requirement. The option "cfm30" selects fast mode (using 5x memory) and a block size of 30 MB. In all experiments, the block size is set larger than the input size.

ppmonstr variant J is the top ranked PPM compressor. It predicts characters one at a time based on the previous 32 bytes (with -o32 option), dropping to a lower order context when no previous match is found. -m1600 selects 1.6 GB of memory to store statistics. When memory is used up, some of the statistics are discarded to make room. Using a lower order conserves memory and improves compression in this case. Otherwise the highest order possible should be used.

Nanozip 0.09a with option -cc and the various PAQ compressors such as paq8pxd and paq8px v69 use context mixing algorithms. Bits are predicted one at a time and arithmetic coded. In the PAQ variants, there are hundreds of models whose predictions are adaptively averaged together, making the programs extremely slow (about 20-30 MB per hour) and memory hungry. Nanozip uses fewer models for better speed. Options

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

select 1.6 GB memory. Statistics are stored in hash tables, discarding old data as they fill up.

Source Code Complexity

To estimate the complexity of a line of code, we compress 29.9 MB of C source code from gimp v2.0.0 (2004), a graphics editor (from the now defunct UCLC compression benchmark), and header files from mingw 4.5.0 (2010), a C++ compiler. The code is as follows:

- gimp *.c, 999 files, 18.180 MB
- gimp *.h, 775 files, 2.414 MB
- mingw *.h, 657 files, 9.299 MB

The total is 927,913 lines of C and C++ code with an average length of 32.2 bytes per line. Compressed sizes are as follows:

```
29,893,907 uncompressed
 5,066,421 zip -9
 3,457,344 7zip -mx
 3,433,685 bbb cfm30
 2,458,090 nanozip -cc -m1600m
 2,450,077 ppmonstr -o32 -m1600
 2,113,906 paq8pxd_v1 -8
 1,919,756 paq8px_v69 -8
 1,865,080 paq8pxd_v4 -8
```

The best result is by paq8pxd_v4, which yields 2.010 bytes or 16.08 bits per line of code.

Human Genome Complexity

The hg19 human reference genome is a consensus of several anonymous humans. It consists of the following files in FASTA format, with sizes shown:

```
254,235,640 chr1.fa
  108,584 chr1_gl000191_random.fa
  558,468 chr1_gl000192_random.fa
248,063,367 chr2.fa
201,982,885 chr3.fa
```

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
194,977,368 chr4.fa
  602,251 chr4_ctg9_hap1.fa
  193,607 chr4_gl000193_random.fa
  195,321 chr4_gl000194_random.fa
184,533,572 chr5.fa
174,537,375 chr6.fa
 4,714,751 chr6_apd_hap1.fa
 4,891,294 chr6_cox_hap2.fa
 4,702,619 chr6_dbb_hap3.fa
 4,776,945 chr6_mann_hap4.fa
 4,930,081 chr6_mcf_hap5.fa
 4,704,239 chr6_qbl_hap6.fa
 5,027,155 chr6_ssto_hap7.fa
162,321,443 chr7.fa
  186,576 chr7_gl000195_random.fa
149,291,309 chr8.fa
   39,715 chr8_gl000196_random.fa
   37,941 chr8_gl000197_random.fa
144,037,706 chr9.fa
   91,909 chr9_gl000198_random.fa
  173,294 chr9_gl000199_random.fa
  190,798 chr9_gl000200_random.fa
   36,893 chr9_gl000201_random.fa
138,245,449 chr10.fa
137,706,654 chr11.fa
   40,929 chr11_gl000202_random.fa
136,528,940 chr12.fa
117,473,283 chr13.fa
109,496,538 chr14.fa
104,582,027 chr15.fa
 92,161,856 chr16.fa
 82,819,122 chr17.fa
 1,714,462 chr17_ctg5_hap1.fa
   38,271 chr17_gl000203_random.fa
   82,960 chr17_gl000204_random.fa
  178,103 chr17_gl000205_random.fa
   41,845 chr17_gl000206_random.fa
 79,638,800 chr18.fa
   4,371 chr18_gl000207_random.fa
 60,311,570 chr19.fa
   94,566 chr19_gl000208_random.fa
```

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
 162,376 chr19_gl000209_random.fa
64,286,038 chr20.fa
49,092,500 chr21.fa
 28,259 chr21_gl000210_random.fa
52,330,665 chr22.fa
 16,909 chrM.fa
 169,914 chrUn_gl000211.fa
 190,612 chrUn_gl000212.fa
 167,540 chrUn_gl000213.fa
 140,489 chrUn_gl000214.fa
 176,012 chrUn_gl000215.fa
 175,756 chrUn_gl000216.fa
 175,608 chrUn_gl000217.fa
 164,386 chrUn_gl000218.fa
 182,798 chrUn_gl000219.fa
 165,055 chrUn_gl000220.fa
 158,521 chrUn_gl000221.fa
 190,615 chrUn_gl000222.fa
 184,081 chrUn_gl000223.fa
 183,303 chrUn_gl000224.fa
 215,413 chrUn_gl000225.fa
 15,325 chrUn_gl000226.fa
 130,958 chrUn_gl000227.fa
 131,719 chrUn_gl000228.fa
 20,328 chrUn_gl000229.fa
 44,581 chrUn_gl000230.fa
 27,950 chrUn_gl000231.fa
 41,482 chrUn_gl000232.fa
 46,876 chrUn_gl000233.fa
 41,358 chrUn_gl000234.fa
 35,180 chrUn_gl000235.fa
 42,789 chrUn_gl000236.fa
 46,801 chrUn_gl000237.fa
 40,754 chrUn_gl000238.fa
 34,517 chrUn_gl000239.fa
 42,788 chrUn_gl000240.fa
 43,012 chrUn_gl000241.fa
 44,410 chrUn_gl000242.fa
 44,224 chrUn_gl000243.fa
 40,744 chrUn_gl000244.fa
 37,401 chrUn_gl000245.fa
```

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

```
  38,934 chrUn_gl000246.fa
  37,167 chrUn_gl000247.fa
  40,598 chrUn_gl000248.fa
  39,289 chrUn_gl000249.fa
158,375,978 chrX.fa
 60,561,044 chrY.fa
3,199,905,909 bytes
```

The files are in FASTA format with a one line header like ">chr1" denoting the file name, followed by lines of 50 bases (A,C,G,T,N) terminated by a linefeed. It uses lowercase letters (a,c,g,t) to indicate tandem repeats. It uses N to indicate unknown bases. These usually occur in large blocks around the centromere (about 40% into most of the large files) and smaller blocks scattered throughout the file and at the telomeres on the ends. Out of a total of 3,137,161,264 bases, 239,850,802 (7.6%) are N.

Unreadable bases typically occur in highly repetitive sections of the code. During shotgun sequencing, the chromosome is broken up into small fragments and sequenced in overlapping "reads" of about 100 bases and reassembled. In repetitive regions, there are multiple ways to reassemble the fragments, making them difficult to sequence. The centromere is the "handle" used to pull apart the two copies of the chromosome during mitosis or cell division. The telomeres are trimmed with each replication to prevent runaway growth.

The files *chr1.fa* through *chr22.fa* are the 22 normal chromosomes. Every cell in the body has two of these, one inherited from each parent. *chrX.fa* and *chrY.fa* are the sex chromosomes. Males have one X and one Y. Females have two X. *chrM.fa* is the mitochondria chromosome, which has its own (slightly different) genetic code. The files ending in *random.fa* are fragments that could not be matched to the main chromosome, so their location is unknown. The files starting with *chrUn* are fragments in which the original chromosome is not known. The files *chr4_ctg9_hap1.fa* and the 7 files *chr6_*_hap?.fa* are small regions of chromosomes 4 and 6 (in the middle of the short arm of 6) that too variable between individuals to form a consensus. These nevertheless have a high degree of overlap.

Only about 1.5% of the human genome consists of exomes, or genes encoding protein. Over half consists of repeating sequences. Some of this serves to regulate genes by binding to proteins that initiate or inhibit transcription. Other sections contain code that is no longer used, or that was inserted by retroviruses and passed on to succeeding generations. Not all of the code is understood.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

There are approximately 20,000 genes in the human genome, although the exact number is not known. In contrast, the 1 millimeter long, bacteria eating roundworm, *C. elegans* has 20,470 protein encoding genes and another 16,000 RNA encoding genes in only 100M base pairs, 3% of the size of the human genome. If we are to believe that humans are more complex than roundworms, then that complexity must somehow be encoded in the "junk" DNA.

The major source of redundancy in the genome (that we know of) comes from repetitive sequences. There may be many adjacent copies, or they may be widely separated or on different chromosomes. They may be on complementary strands. Only one strand on each chromosome is recorded. The opposite is formed by matching A to T and C to G and reversing the order, for example, *TACT -> AGTA*. None of the compressors in our test set are able to recognize complementary strands as contexts. Also, because of the large size of the genome, none of the compressors is able to recognize long distance matches except for *BBB*, and then only if the genome is represented in a more compact form than one base per character.

The obvious way to pack DNA is to use 2 bits per base and 4 bases per byte. However this can make compression worse because two identical strings will appear different to the compressor unless the distance between them is a multiple of 4. To solve this problem, we pack 3 or 4 bases into a byte such that after a while the byte boundaries synchronize. The code we use is the same as the FASTQZ compressor (Bonfield and Mahoney, 2013). The bases A, T, C, G are encoded as 1, 2, 3, 4 respectively and grouped such that when interpreted in base 4, they form a number in the range 64 to 255. This means that any group starting with G, CG, or CCG is packed 3 to a byte, and all others 4 to a byte. The following example shows how bases would be grouped using different starting points.

TGGA ATCA GAT GGA ATCA TCGA ATGG ACTG GAA TGGA ATCA
GGA ATCA GAT GGA ATCA TCGA ATGG ACTG GAA TGGA ATCA
GAAT CAGA TGGA ATCA TCGA ATGG ACTG GAA TGGA ATCA
AATC AGAT GGA ATCA TCGA ATGG ACTG GAA TGGA ATCA

We give higher codes to C and G because they occur less frequently than A and T in the human genome, resulting in tighter packing before compression. Also, we discard all N, under the assumption that the data is highly repetitive and therefore contains very little information. We then compress two ways, once as a single file and once as 26 files. The 26 files are chromosomes 1 through 23, X, Y, M, and Unknown, formed by removing N, concatenating the remaining bases, and packing as described.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Variants (chromosomes 4 and 6) and random fragments are concatenated to the chromosomes to which they belong. The unknown fragments go in their own file. For the single file, the compressed sizes (in bytes) are as follows:

766,373,649 uncompressed
683,485,287 zip
622,113,887 7zip
605,526,316 bbb
599,775,019 ppmonstr -o8
598,722,820 nanozip

As 26 separate files the total compressed sizes are as follows:

766,373,636 uncompressed
683,494,823 zip
630,447,231 7zip
628,334,542 bbb
615,908,412 ppmonstr -o8
611,444,955 nanozip
604,332,601 paq8pxd_v1

Compression with *paq8pxd_v1* took 40.6 hours on a 2.0 GHz T3200 processor. Sampling a few files with *paq8pxd_v4* showed that compression would have been worse in spite of being a newer version. The better compression on the source code was due to fixing a problem with overly aggressive file segmentation, which was not a problem with the DNA.

The difference in compressed size for *BBB*, 22,808,226 bytes, is an estimate of the mutual information between chromosomes. The difference is smaller for all other compressors because they could not store the complete statistical model in the 1600 MB of available memory. (*BBB* stores the model in 766 MB of memory and 3 GB of temporary files for the suffix array). This suggests that *paq8pxd_v1* would have compressed to 581.5 MB given sufficient memory.

To test the effects of including the variants of *chr6*, we compared the compressed sizes (after packing) of *chr6.fa* alone and with the 7 variants concatenated onto the end. The results show that appending the variants only has a very small effect on the total information content, adding 0.26 MB using the best compressor tested.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

chr6 only plus variants
44,180,099 51,553,182 packed only
36,762,504 38,590,394 bbb
36,338,669 36,694,017 ppmonstr -o8
36,113,475 36,370,417 nanozip

Although we packed bases with a self-synchronizing code, we nevertheless lose some compression at the beginning of the match before synchronization. To test this effect, we compare compression with and without packing of *chr21* and *chr22*. The unpacked input contains only the letters A, C, G, T, converted to uppercase, discarding the FASTA header, newlines, and all N.

chr21 Packed Chr22 Packed
35,106,642 9,287,838 34,894,545 9,341,723 Uncompressed
7,884,270 7,949,255 7,538,934 7,580,025 bbb
8,134,163 7,802,206 7,853,008 7,377,597 ppmonstr -o8
7,906,238 7,744,100 7,482,284 7,308,297 nanozip

For *nanozip* and *ppmonstr*, packing improves compression because it reduces memory usage and effectively increases the context order. For *BBB*, compression is 0.82% worse for *chr21* and 0.55% for *chr22*. *BBB* (BWT) uses an unbounded context order and is not limited by memory. This suggests that compression could be improved by 2 or 3 MB overall (about 579 MB) by not packing if sufficient memory were available.

Finally, to test the effects of reverse complement contexts, we compute the reverse, the complement, and the reverse complement of *chr21* and *chr22* and append these to the original data (unpacked, but reduced to A, C, G, T as before). The reduction in size of the reverse complement over the other two gives us an estimate of the space that could be saved by recognizing such contexts. Results using *BBB* are as follows:

15,768,540 chr21 x 2
15,951,005 chr21 + complement
15,950,060 chr21 + reverse
15,633,621 chr21 + reverse complement

15,077,868 chr22 x 2
15,288,313 chr22 + complement
15,287,298 chr22 + reverse
14,895,566 chr22 + reverse complement

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Appending the reverse or the complement makes compression worse than storing two compressed copies of the original file. But appending the reverse complement improves compression by 0.86% for *chr21* and 1.20% for *chr22*. This suggests that a 1% (6 MB) improvement might be possible overall, for an information content of 573 MB or 4.58 x 10^9 bits.

There may be many other sources of redundancy that might be discovered with improved compression techniques. That is an area of future work.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART IV:

NEUROSCIENCE AND PHILOSOPHY OF MIND

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER ELEVEN

HYPOTHESES ABOUT THE INTEGRATION OF CORTICAL ACTIVITY: PSYCHOLOGICAL AND PHYSIOLOGICAL 'BINDING'

ALFREDO PEREIRA JR.

In this paper I discuss hypotheses about cortical integration in the recent history of neuroscience, and possible correlations between cortical processes and the psychological phenomenon of perceptual integration. The discussion of both – physiological and psychological processes – is usually referred to as the "binding problem". Although proposing that both dimensions are closely inter-related, I criticize the conflation of both approaches. Concerning physiological integration, I propose an analogic model that combines subcortical control of cortical activity with mechanisms intrinsic to the cortical tissue.

Introduction

The distributed architecture of the neocortex has inspired various attempts to explain how the functions of the parts are integrated. The discovery of specialization of function, beginning with the studies made by Broca in 1876, has revealed the existence of regions sensitive to signals from receptors in each sensory modality, as well as "associative" areas where the sensory regions converge. Besides such specialization at the macro level, there is also segmentation at the micro level. "Columns" with millions of neurons in an approximate area of 1 square millimeter, and (on average) six "layers" can be identified in the neocortical tissue. Mountcastle (1979) proposed a model of vertical columns and horizontal layers, forming functionally specialized micro "modules". Such use of the term cannot be conflated with the contemporary one (inspired by Fodor,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

1983) that refers to macro brain modules supposed to be responsible for mental functions.

In the recent history of neuroscience, two kinds of models of cortical integration may be distinguished: those that propose the integration to be made autonomously by the cortical network (e.g. Burnod, 1990; Pribram, 1991) and those that emphasize the control exerted by subcortical structures (the most famous being Crick, 1983, 1994). In fact, all neocortical subsystems are interconnected by subcortical structures; the thalamus in particular may be considered an omnipresent coordinator underlying cortical function. The main difference between the two models lies on the character of the subcortical assistance to the neocortex, and the corresponding degree of autonomy of intra and inter-cortical connections and cognitive processing.

For the first current – here called "the tangential view" – the subcortical role is auxiliary, in terms of releasing transmitters that activate cortical synapses, inducing collective oscillations that "carry" sensory and cortically generated patterns. In this view, the intrinsic cortical connections and processes are presumed to be sufficient to account for the functional integration that supports cognitive processing. For the second current – "the radial view" – the flux of information processed in the cortex goes back and forth to the thalamus and other subcortical structures that are presumed to control cortical dynamics (e.g., synchronous collective oscillations) analogously to the way a puppet is controlled by the hands of its manipulator. In the radial view, intra and inter-cortical connections are assumed to have only a modulatory role for cognitive processing (e.g., see Phillips and Singer, 1997, on the concept of "contextual fields").

The Historical Conflict between Views of Cortical Integration

The first general hypothesis of cortical integration was the theory of a tangential electromagnetic field. Two pioneers of electroencephalography, R. Gerard and B. Libet (Gerard and Libet, 1940), proposed that cortical activity is integrated by an electromagnetic field tangential to the scalp. The scalp EEG was considered a measure of such field. The theory was widely defended by W. Kohler, who saw in it a neurophysiological basis for his psychological theory of a perceptual field. Kohler also made experiments to prove the theory (Kohler and Held, 1949; Kohler, Held and O'Connell, 1952).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, the acceptance of the lateral field theory was soon blocked by a series of experiments made by R. Sperry, who attempted to prove its falsity. He made tangential cuts in the cortex of rats, showing that after the lesions the animals were still able to perform associative tasks and spatial orientation (Sperry, 1955; see criticisms in Kohler, 1965). Contrary to the interpretation that was given to the results at that moment, we can show today that they were not conclusive about the absence of lateral connections in the cortex. For example, lateral propagation by diffusion of chemicals would be expected to occur in the scale of less than two millimeters, while Sperry's lesions were not so accurate. The insulation should be proportional to this size in order to produce major behavioral deficits. The cortico-cortical connections by bundles of axons should be cut to certify that there is no horizontal transmission. Functions that do not depend crucially on the neocortex, as spatial orientation, cannot be impaired by such lesion and insulation procedures. Behavioral performances should be accurately described, because we know that some simple reflexes or even simple associations don't need much inter-(neo)cortical processing. However, the effect of Sperry's weakly supported conclusions was so influential that more than 40 years later neuroscientists were surprised with the "discovery" of horizontal connections in the visual, auditory and prefrontal cortex (Gilbert, 1995).

Discoveries about thalamo-cortical connectivity and columnar specialization, from the fifties to the seventies, led to the current dominant view that the integration of cortical activity is made by the thalamic-reticular system. Integrative cognitive processing is assumed to correspond to electromagnetic processes in the direction radial to the scalp. In fact, there are many bidirectional (afferent and efferent) connections between practically all cortical regions of the cortex and the thalamus. The afferent connections project to cortical layer IV, and the efferent connections come mostly from cortical layers V-VI. Some facts seem to corroborate the radial theory: first, many of the axons of the pyramidal cells are directed downwards, suggesting that the preferential direction for the flow of energy/information would be toward the deeper layers and finally the thalamus. Second, it is well known that the reticular-thalamic system coordinates the rhythms of neuronal firing in the cortex, by means of releasing excitatory and inhibitory transmitters.

However, experimental discoveries and explanatory requirements suggest the possibility of a tangential theory:

a) There are horizontal connections in the visual cortex (Gilbert, 1995) and very impressive bundles of horizontal fibers - called

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

"stripes" - in the prefrontal cortex (Levitt et al., 1993; Ritzer and Goldman-Rakic, 1995; Goldman-Rakic, 1995; Melchitzky et al., 1998);

b) The tangential view affords better explanation of the function of layers I, II and III, i.e., to promote the horizontal flow of the excitatory potential;
c) The tangential view provides good hypothesis about how dopamine and serotonin work in the prefrontal cortex. Such hypothesis fit with the data recently obtained by the role of such transmitters, and their relation to psychiatric phenomena (see e.g. Krimer et al., 1997; Zahrt et al., 1997);
d) It also provides better explanation of magneto encephalographic (MEG) measurements of neuronal activity during cognitive tasks. The MEG measures only fields tangential to the scalp, while the EEG measures both tangential and radial fields. The radial hypothesis is not able to explain MEG measurements convincingly; their defenders simply assume that MEG data is produced by pyramidal neurons located tangentially to the scalp (of course such neurons exist, since cortical folding forms bumps - but neurons in this position are likely to constitute a minority);
e) Last but not least, it is consistent with the intuitive belief that the relatively small thalamus, besides its other known functions, doesn't have computational power to handle the integration between the relatively large cortex.

The debate between both views reemerged in the study of the primary visual cortex. The classical view proposed by Hubel and Wiesel (1962) is that thalamo-cortical pathways determine orientation selectivity. A new paradigm, in some aspects similar to Kohler's, was recently proposed, stating that reentrant signaling through cortico-cortical connections would have the predominant role in such cognitive function (see Pei et al., 1994; Ben-Yishai et al., 1995; Carandini and Ringach, 1997; Ringach et al., 1997, and Ringach, 1998).

The fact that both views have good evidences in their favor, as well as explanatory power, suggests an interesting solution to the debate, in terms of a theoretical model that combines tangential and radial mechanisms. Before presenting a sketch of such model, I will discuss the relation between cortical integration and psychological "binding".

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Integration and "Binding"

Binding phenomena are omnipresent in biological systems: proteins bind to effectors, organisms bind for reproduction, and in human language verbs bind to predicates. In cognitive neuroscience, a similar "binding" problem has become famous, the problem of explaining how sensations from different modalities bind to produce a unified perception of the world (Crick, 1994; Hardcastle, 1994; Treisman, 1996). Unfortunately, although being an exciting discussion, such a formulation of the problem conflates three inter-related but different problems:

a) The problem of cortical integration;
b) The problem of inter-modal integration, that is relatively independent of cortical mechanisms, since in many species (see Stein and Meredith, 1993) inter-modality is performed by the superior colliculus. In humans this subcortical structure has a relatively smaller size, having a more limited function of controlling eye movements (see Schiller, 1997). This fact justifies the association between cortical integration and perceptual binding for humans, but the generalization to non-primates is not adequate;
c) The problem of psychological binding, i.e., how different aspects of perception (e.g., in visual perception, form, color and movement), presumably processed by distributed cortical systems, are unified in a single phenomenal world. Treisman (1996) identified seven different types of perceptual binding:
   - binding of properties (shape, color) to objects;
   - binding of the parts of an object against a background;
   - binding of particular values on a dimension (e.g., ranges of light frequency that are bound into the same color);
   - hierarchical binding, e.g. binding of boundaries to surfaces;
   - conditional binding, where "the interpretation of one property (e.g. direction of motion) often depends on another (e.g. depth, occlusion or transparency)" (p.171);
   - temporal binding of successive states of the same object;
   - location binding of objects to their current locations.

The formulation of the "binding" problem in cognitive neuroscience usually conflates physiological and psychological concepts. For example, Gegenfurtner (1998) says "human vision is commonly divided into low level and high level processes. At the lower level, feature analysis occurs to extract information about edges, colors or depths at a particular location

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

in the visual field. At a higher level, the visual system must recognize the objects that are built up from these low feature levels. But there is also an intermediate stage, which presents a major challenge to any theory of vision: how does the visual system know which features belong to which object? ...In order to link features into objects, signals from different parts of the retina, mapped onto different parts of the retinotopic cortical area, must be bound together" (p. 96). Such a conflation of problems reveals that cognitive neuroscientists are still struggling with the theoretical relation between sensation (the neurophysiological effects of the presentation of a stimulus, or "low level" visual processes) and perception (the mental or subjective interpretation of the sensation, or "high level" vision).

In the first half of the century, psychologists used to completely separate sensation from perception; sensation was considered the measurable effect of stimulation upon the nervous system, and perception was relative to the non-measurable phenomenal experiences elicited by the stimulation. Psychophysics was then considered a study of the measurable correlations between properties of the stimulus and sensations, leaving the study of perception to philosophers and psychologists. However, some of the Gestaltists were already trying to establish correlations (or even a stronger "isomorphism") between sensory phenomena defined in terms of neuronal sensibility and perceptual phenomena like aftereffects and figure/ground segregation.

Teuber (1960), following a theoretical discussion made by Boring, consistently argued for the continuity between sensation and perception. If such continuity is assumed, the solution of the neurophysiological problem of integration should help to solve the psychological problem of "binding". For example, in 'gestalt' psychology perceptual fields are assumed to be supported by cortical electromagnetic fields. Nevertheless, in the contemporary discussion of "binding" reductionist assumptions have pushed the conceptual framework to the other extreme, i.e. the complete identification between neuronal mechanisms of sensation and perceptual phenomena, with two consequences:

a) Careless imports from neuroscience to psychology. Physiological mechanisms are automatically taken as psychological. The existence of single neurons or small cell assemblies that work as "feature-detectors" is assumed to imply that psychological processes are processed in completely independent parallel streams to be "bound" together at some point. This view is illustrated in Gegenfurtner's (op.cit.) metaphor of binding occurring in an "intermediate level";

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

b) Careless imports from psychology to neuroscience. The apparent absence of an explication of the mechanisms by which different aspects of perception are bound together is identified with the problem of how neurons in different parts of the cortex communicate with each other. In this view, the unsolved psychological phenomenon of binding is assumed to imply the existence of an unsolved problem of cortical integration. Although there must be a strong relation between physiological and psychological binding, such identification is misleading. Neuroscientists have known for a century the existence of widespread intra and inter-cortical connections, including convergent pathways to "associative" areas. What remains to be explained is how such connections and pathways support psychological binding.

The most plausible view, according to the idea of a complementarity of the radial and tangential models, would be that primary and associative perceptual areas form an integrated network, combining forward and feedback connections, simultaneously processing features of the stimuli and the respective integrated percept. There are currently three hypotheses consistent with this approach, that I call the "synchrony", the "spatio-temporal coherence" and "resonance" hypotheses. All of them are based on electrical patterns of neuronal activity, measured by EEG or arrays of invasive electrodes in experimental animals. Influential neuroscientists working on the "binding" problem have assumed the first one (see Treisman, 1996), but as synchrony entails informational redundancy, it is difficult to understand how synchronous activity would support a variety of conscious experiences. Spatio-temporal coherence (see e.g. Roy John et al., 1997) is an interesting and powerful hypothesis, which includes and goes beyond synchrony. One possibility of generating coherence is the temporal autocorrelation of electric patterns. There are good neurophysiological evidences that temporally structured patterns correlate with the experience of 'qualia' (an excellent article, that reviews such evidences, is Cariani, 1994). The resonance hypothesis goes one step further, accounting for processes of "reciprocal causality" (when two neurons or two assemblies resonate, each one amplifies the activity of the other). If reciprocal causality has a special role in physics (e.g., non-linearity), then the resonance hypothesis may have an explanatory role in neuroscience (it has already proven to be powerful in artificial network modeling).

The discussion of the three hypotheses above requires the sharpening of conceptual definitions. For example, in an influential article on the "binding" problem Valerie Hardcastle (1994) conflates oscillatory

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

synchrony and resonance between cortical columns. In fact, the concept of inter-columnar resonance is able to account for the cognitively relevant kinds of firing synchrony (Grossberg and Somers, 1992), but the converse is not true. Oscillatory synchrony may be produced by sub-cortical mechanisms without any direct causal relation between the cortical neurons that synchronize; resonance, on the other hand, implies reciprocal causality between cortical neurons.

Coherent (including resonating) electric patterns of neuronal activity are a part of a larger causal chain. The patterns are directly controlled by biochemical processes at the synapses (transmitters coming from the axon of the pre-synaptic neuron, receptors produced in the post-synaptic neuron, calcium from the surrounding environment. etc.). At the same time, afferent patterns influence the morphology and function of perceptual neurons (and, probably, indirectly influence all neurons). The presence of two classes of factors, internal and external to the brain, generating coherent activity of neurons, suggests that "binding" should be supported by a complex neuronal mechanism.

The adoption of any of the above hypotheses is not at this moment sufficient to explain psychological/phenomenological evidences. One of the reasons of this difficulty is that at the conscious level we are always in the presence of an integrated and unitary phenomenal world; we cannot distinguish between perceptual objects before being integrated and after being integrated. There is no report of cases of neuronal tissue lesion where the subjects fail to perform "binding". The reported cases (see Treisman, 1996) show patients that are not able to perform one putative step of the "binding" process, or are not able to perceive some type of stimulus or location at all. However, the remaining perceptual capabilities of these patients always preserve diverse – if not all – modalities of "binding". It seems to be an essential part of the continuous process that goes from sensation to perception and not a separate intermediate stage "between" sensation and perception. In this sense, any elementary sensation that is consciously perceived already displays some modality of "binding".

The bad consequence for scientific research would be that we cannot introspectively identify the steps of binding processes, and therefore such steps cannot be experimentally correlated with specific neuronal processes. Such a skeptic view contradicts Treisman's optimistic statement that "the strongest evidence will come when changes in neural activity are found to coincide with perceived changes in binding, perhaps in ambiguous figures or in attentional capture" (Treisman, 1996). If there is not such a thing as "perceived changes in binding", even when neuroscientists come to discover all mechanisms of neuronal integration, a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

problem will remain, about which of them support unconscious psychological binding. Given such a methodological limitation, I will return to the discussion of cortical integration, proposing a model that is not able to solve the "binding" problem but has other useful applications.

A Synthesis of the Radial and Tangential Models

I propose a synthesis of the tangential and radial models of cortical integration, designed to show how the excitatory potential in a column (corresponding to a resonating patterns) could directly propagate to other columns, induced by oscillatory collective activity controlled by the thalamus and basal ganglia. The horizontal flow of the excitatory potential may have two modalities: horizontal axonal and dendrodendritic (see Shepherd et al., 1985) transmission, and diffusion processes (including the diffusion of transmitters, diffusion of calcium transported by glial cells, etc.). The combination of both forms gives to the horizontal flow some deterministic as well as some random aspects.

An analogic model of the dynamics of pyramidal neurons of the prefrontal cortex is presented below, describing how sequential cognitive processing could be supported by cortical integration. Subcortical structures are assumed to have a necessary function for the cross-cortical flow of energy and information. The 30-50 Hz oscillations, induced by reticular-thalamic transmitter release, are proposed to combine with inhibition of apical dendrites in intermediary and deeper layers of cortical columns (e.g., dopaminergic inhibition in the pre-frontal cortex) to produce a horizontal flow of the excitatory potential. Such horizontal flow would be able to activate resonating columns sequentially, even in the absence of new input.

The model combines two kinds of oscillatory mechanism:

a) Arousal-Attentive Reticular-Thalamic Oscillations (see e.g. Llinas and Ribary, 1992, 1994). The oscillations display Beta frequency (30-50 Hz), being regulated by two main transmitters: Glutamate (excitatory) and GABA (inhibitory);
b) Inhibition-Disinhibition of Apical Dendrites, studied by Glowinski et al. (1984), and recently by Yang and Seamans (1996), and Krimer, Jakab and Goldman-Rakic (1997). Apical dendrites in the prefrontal cortex are alternately excited and inhibited. The action of diverse dopamine and serotonin receptors would be related to inhibition/disinhibition of apical dendrites.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

One possible function of this mechanism would be the control of energy/information flow between layers in the following way: the blocking of radial flow between layers in a column promotes tangential flow (lateral propagation), and the excitation of spines disinhibiting the apical dendrites promotes radial flow (reentrant loop) between layers of the column. This hypothesis is consistent with neuropsychological studies that suggest a role for the basal ganglia in high cognitive processing (e.g., Hayes et al., 1998). It would be a neurobiological basis for Burnod's theory of functional networks between cortical columns, named "call trees" (Burnod, 1990).

The model uses the following conventions. The possible states of neurons are classified in three categories: a) the neuron is firing; b) sub-threshold excitatory potential; c) the neuron is inhibited by external agents. The six cortical layers are classified in three categories: input, output and horizontal connections. The "input" layer corresponds to cortical layer 4, for thalamic input, and deep layer 3, which is assumed to have receptive fields for signals coming from parietal and temporal "associative" areas, as well as other cortical sources. The incoming information is assumed to be encoded in temporal patterns of firing (interspike intervals, temporal coincidences, etc.; see Cariani, 1994), being broadcasted to a large number of columns.

The column(s) that respond to the temporal pattern of the stimulus – possibly on the basis of the structure of the dendritic tree, and transmitter/receptor distributions determined by previous learning – are assumed to be the input columns. In the model, for the sake of simplification, only four columns are depicted (numbered 1 to 4), and only one input is considered, but of course this basic mechanism is assumed to occur repeatedly in space and time. The "output" layer in the model (encompassing cortical layers 5 and 6) is understood as the layer that sends the results of prefrontal processing to other parts of the brain ("associative" areas, pre-motor or motor cortex, etc.). The output columns are determined by the trajectory of the cortical processing. The "horizontal connection" layer corresponds to cortical layers 1, 2 and superficial 3, and is assumed to be activated mainly by feedback from the output layer, and to send signals for deep 3 layer in other cortical columns.

The sequential mechanism is conceived in four steps, following the temporal dynamics induced by reticular/thalamic oscillations:

Step 1: t = 0 to 12 ms, excitatory plateau
- Layer IV and deep III pyramidal cells respond to input signal

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Step 2: t = 13 to 24 ms, inhibitory valley
- Neurons of the output layer of column 2 fire, defining the first item of a serial order
- Feedback to Layers I-II-supIII of column 2
- Inhibition of apical dendrites of Layer IV and deep III neurons in col. 2
- Lateral inhibition of col. 2

Step 3: t = 25 to 37 ms, excitatory plateau
- Lateral excitation of column 2
- Horizontal flow of excitatory potential in Layers I-II-supIII, reaching input layer of column 4.

Step 4: t= 38 to 50 ms, inhibitory valley
- Lateral inhibition of column 4
- Firing of output neurons in col. 4, corresponding to the second item of the series
- Feedback to Layers I-II-supIII of col. 4
- Inhibition of apical dendrites of pyramidal cells in col. 2

The model describes how Reticular-Thalamic 30-50 Hz oscillation, combined with inhibition of apical dendrites of pyramidal cells of the intermediary layers of the input column, could produce a horizontal flow of the excitatory potential, sequentially activating other columns. The trajectory of the horizontal flow, sequentially activating different output columns, defines the items of a serial order. The sequential activation occurs in the absence of new stimuli, fulfilling the requirements for an internally generated serial behavior.

Such mechanism probably works together with reception of new stimuli, resulting in a complex interplay of internally generated and externally received patterns. Realistically, this kind of mechanism is able to generate multiple parallel streams, some of them conscious (e.g., semantic processing) and some not conscious (e.g., processing of the sequence of gestures during speech). The prefrontal cortex seems to be the best candidate for the production of linear sequences, while other parts of the brain would be more specialized for the coordination between the items of different sequences (e.g., the posterior parietal cortex for visuomotor coordination, the hippocampus for coordinating egocentric and allocentric spatial sequences, the cerebellum for monitoring the execution of diverse serial commands from the motor cortex, etc.).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Applications of the Model

There are many cognitive functions that putatively depend on cortical integration. First, automatic (primed/conditioned) associations can be distinguished from non-automatic (those that require logical inference), in terms of the cortical mechanisms that are involved in each case. In the case of a primed stimulus, the columns that recently responded to it have available receptors that facilitate them to be the sink of the excitatory potential. In the case of strongly conditioned associations, it is possible that neuronal connections were developed between the columns that respond to the stimulus and the columns that trigger the response. However, in the case of reasoned associations, a larger horizontal activity is required, corresponding to possible inhibition of primed/learned associations and search for an adequate response. This hypothesis, which follows from the model, is consistent with the observation of a stronger negative component in EEGs, in tasks that require sequential reasoning (Kutas and Hillyard, 1980).

An experimental paradigm for the dissociation of automatic and non-automatic linguistic responses (Kane, Picton, Moscovitch and Winocur, 1998) uses the following kind of procedure: a word is primed (e.g., "ingenuous") and a sentence to be completed is presented (e.g., "Einstein was an _____ man"). The cases when the subjects complete the sentence with the primed word ("ingenuous") is likely to produce a prefrontal negative component that lasts longer than when the subject inhibits the primed word and rehearsals a semantically more appropriate word ("ingenious"). According to the model presented here, in the "automatic" case the horizontal flow was soon interrupted, because the excitatory potential readily found a sink (the previously primed column); in the reasoned case, the horizontal flow has longer duration, corresponding to inhibition of the primed column and search for the semantically correct one (OBS.: this effect depends on the position of the scalp electrodes).

Another use of the model would be making sense of the mechanism underlying tangential dipoles measured by MEG, while human subjects are performing cognitive tasks. The hypothesis would refer to a summation of components of horizontal flow over intervals of hundreds of milliseconds, generating the dipoles measured by MEG. This possibility is not trivial, since the current opinion about the sources measured by MEG is based on the radial model, and consequently tends to consider only the activity of pyramidal cells parallel to the scalp. However, the proportion of cells localized in this direction is probably too small to account for the magnitude of MEG measurements.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Acknowledgment

FAPESP for support of my Post-Doctoral period, when this work was done, and also for support of my current research.

References

Ben-Yishai, R., Bar-Or, R. L. & Sompolinski, H. (1995) Theory of Orientation Tuning in Visual Cortex. *Proceedings of the National Academy of Sciences USA* **92** (9): pp. 3844-3848.

Bressler, S.L. (1995) Large Scale Neuronal Networks and Cognition. *Brain Research Reviews* **20**, pp. 288-304.

Burnod, Y. (1991) An Adaptive Neural Network: The Cerebral Cortex. London: Prentice Hall.

Carandini, M. & Ringach, D. L. (1997) Predictions of a Recurrent Model of Orientation Selectivity. *Vision Research* **37** (21): pp. 3061-3071.

Cariani, P. (1994) As Time Really Mattered: Temporal Strategies for Neural Coding of Sensory Information. In Pribram, K. (ed.) *Origins: Brain and Self-Organization*, Hillsdale: Erlbaum.

Cauller, L. J. & Kulics, A. T. (1991) The Neural Basis of the Behaviorally Relevant N1 Component of the Somatosensory-Evoked Potential in SI Cortex of Awake Monkeys: Evidence That Backward Cortical Projections Signal Conscious Touch Sensation. In: *Experimental Brain Research* **84**, pp. 607-619.

Crick, F. (1984) Function of the Thalamic Reticular Complex: The Searchlight Hypothesis. In: *Proceedings of the National Academy of Sciences USA* **81**, pp. 4586-4590.

—. (1994) *The Astonishing Hypothesis*. New York: Charles Scribner's/Maxwell McMillan.

Crick, F. & Koch, C. (1995) Are We Aware of Neural Activity in Primary Visual Cortex? In: *Nature* **375**, pp. 121-123.

Damásio, A. R. & Damásio, H. (1994) Cortical Systems for Retrieval of Concrete Knowledge: the Convergence Zone Framework. In: C. E. Koch and J. L. Davis (Eds.) *Large-Scale Neuronal Theories of the Brain*. Cambridge: MIT Press.

D'Esposito, M. & Grossman, M. (1996) The Physiological Basis of Executive Functions and Working Memory. *The Neuroscientist* **2**, pp. 345-352.

Duncan, J. (1995) Attention, Intelligence and the Frontal Lobes. In Gazzaniga, M. (1995) *The Cognitive Neurosciences*. Cambridge: MIT Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Fodor, J. (1976) *The Language of Thought.* Cambridge: MIT Press.
—. (1983) *The Modularity of Mind*. Cambridge: MIT Press.
Freeman, W.H. & Schneider, W. (1982) Change in Spatial Patterns of Rabbit Olfactory EEG with Conditioning to Odors. *Psychophysiology* 19, pp. 44-56.
Frith, C. D. & Friston, K. J. (1997) Studying Brain Function With Neuroimaging. In Rugg, M.C. (Ed.) *Cognitive Neuroscience*. Cambridge, MIT Press.
Gallese, V., Fadiga, L., Fogassi, L. & Rizzolati, G. (1996) Action Recognition in the Premotor Cortex. *Brain* **119**, pp. 593-609.
Gazzaniga, M. S. (ed.) (1993) *The Cognitive Neurosciences*. Cambridge: MIT Press.
Gegenfurtner, K. (1998) Visual Psychophysics: Synchrony in Motion. *Nature Neuroscience* **1** (2), pp. 96-99.
Gilbert, C. D. (1995) Dynamic Properties of Adult Visual Cortex. In Gazzaniga, M. (Ed.) *The Cognitive Neurosciences*. Cambridge: MIT Press.
Glowinski, J., Tassin, J. P., & Thierry, A. M. (1984) The Mesocortico-prefrontal Dopaminergic Neurons. *Trends in Neuroscience* **7**: pp. 415-418.
Goldberg, E. (1990) Higher Cortical Functions in Humans: The Gradiental Approach. In E. Goldberg (Ed.) *Contemporary Neuropsychology and the Legacy of Luria*. Hillsdale: Lawrence Erlbaum.
Goldman-Rakic, P. S. (1987) Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. In V. Mountcastle (Ed.) *Handbook of Physiology*, Section 1: The Nervous System, Volume V: Higher Functions of the Brain, Part 1. Bethesda: American Physiological Society.
Goldman-Rakic, P. S. (1996) Cellular Basis of Working Memory. *Neuron* 14: pp. 477-485.
Goodale, M. & Milner, A. (1992) Separate Visual Pathways for Perception and Action. *Trends in Neuroscience*, l5, pp. 20-25.
Goodale, M. (1997) Pointing the Way to a Unified Theory of Action and Perception. *Behavioral and Brain Sciences* **20** (4): pp. 749-750.
Gray, C. (1994) Synchronous Oscillations in Neuronal Systems: Mechanisms and Functions. *Journal of Computational Neuroscience* **1**, pp. 11-38.
Gray, C. & Singer, W. (1989) Stimulus-Specific Neuronal Oscillations in Orientation Columns of Cat Visual Cortex. *Proceedings of the National Academy of Sciences* USA 86, pp. 1698-1702.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Grossberg, S. & Grunewald, A. (1997) Cortical Synchronization and Perceptual Framing. *Journal of Cognitive Neuroscience*, 9, pp. 117-132.

Grossberg, S. & Somers, D. (1992) Synchronized Oscillations for Binding Spatially Distributed Feature Codes into Coherent Spatial Patterns. In Carpenter, G. and Grossberg, S. (Eds.) *Neural Networks for Vision and Image Processing*. Cambridge: MIT PRess.

Grossberg, S., Mingolla, E. & Ross, W. D. (1997) Visual Brain and Visual Perception: How Does the Cortex do Perceptual Grouping? *Trends in Neurosciences*, 20, pp. 106-111.

Gerard, R. W. & Libet, B. (1940) The Control of Normal and "Convulsive" Brain Potentials. *Americal Journal of Psychiatry* 96: pp. 1125-1153.

Hardcastle, V. G. (1994) Psychology's Binding Problem and Possible Neurobiological Solutions. Journal of Consciousness Studies 1 (1): pp. 66-90.

Hayes, A. E., Davidson, M. C., Keele, S. W. & Rafal, R. D. (1998) Toward a Functional Analysis of the Basal Ganglia. *Journal of Cognitive Neuroscience* **10** (2): pp. 178-198.

Hebb, D. (1949) *The Organization of Behavior: A Neurophysiological Theory*. New York: John Wiley.

Hein, A. & Held, R. (1962) A Neural Model for Labile Sensorimotor Coordinations. In *Biological Prototypes and Synthetic Systems*, Vol. 1. New York: Plenum Press.

Held, R. & Hein, A. (1958) Adaptation of Disarranged Hand-Eye Coordination Contingent Upon Re-Afferent Stimulation. *Perceptual and Motor Skills* **8**, pp. 87-90.

Held, R. & Hein, A. (1963) Movement-Produced Stimulation in the Development of Visually Guided Behavior. *Journal of Comparative and Physiological Psychology* **5**, pp. 872-876.

Held, R. (1989) Perception and its Neuronal Mechanisms. *Cognition* **33**, pp. 134-159.

Hubel, D. N. & Wiesel, T. N. (1962) Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. *Journal of Physiology* **160**, p.106-154.

Hubel, D. N. & Wiesel, T. N. (1968) Receptive Fields and Functional Architecture of Monkey Striate Cortex. *Journal of Physiology* **195**, pp. 215-243.

Jonides, J. (1995) Working Memory and Thinking. In E.E. SMITH & D.N. OSHERSON (Eds.) *An Invitation to Cognitive Science*, Vol.3: *Thinking*. Cambridge: MIT Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Jonides, J. & Smith, E. E. (1997) The Architecture of Working Memory. In Rugg, M.D. (Ed.) *Cognitive Neuroscience*. Cambridge, MIT Press.

Kelso, J. A. S. (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge: The MIT Press.

Kohler, W. & Held, R. (1949) The Cortical Correlate of Pattern Vision. *Science* 110: pp. 414-419.

Kohler, W. (1951) Relational Determination in Perception. In L.A. Jeffress (Ed.) *Cerebral Mechanisms in Behavior: The Hixon Symposium*. New York: Hafner.

Kohler, W., Held, R. & O'Connell, D. N. (1952) An Investigation of Cortical Currents. *Proceedings of the American Philosophical Society* **96**, pp. 290-330.

Kohler, W. (1965) Unsolved Problems in Figural Aftereffects. *The Psychological Record* 15: pp. 63-83.

Kohler, S, & Moscovitch, M. (1997) Unconscious Visual Processing in Neuropsychological Syndromes: A Survey of the Literature and Evaluation of Models of Consciousness. In Rugg, M.D. (Ed.) *Cognitive Neuroscience*. Cambridge: MIT Press.

Krimer, L. S., Jakab, R. L. And Goldman-Rakic, P. S. (1997) Quantitative Three-Dimensional Analysis of the Catecholaminergic Innervation of Identified Neurons in the Macaque Prefrontal Cortex. *The Journal of Neuroscience*, 17 (19): pp. 7450-7481.

Kritzer, M. F. And Goldman-Rakic, P. S. (1995) Intrinsic Circuit Organization of the Major Layers and Sublayers of the Dorsolateral Prefrontal Cortex in the Rhesus Monkey. *Journal of Comparative Neurology*, 359: pp. 131-143.

Kutas, M. & Dale, A. (1997) Electrical and Magnetic Readings of Mental Functions. In Rugg, M.D. (Ed.) *Cognitive Neuroscience*. Cambridge, MIT Press.

Lane, R. D., Reiman, E. M., Axelrod, B., Yun, L., Holmes, A. & Schwartz, G. E. (1998) Neural Correlates of Levels of Emotional Awareness: Evidence on an Interaction between Emotion and Attention in the Anterior Cingulate Cortex. *Journal of Cognitive Neuroscience* **10** (4), pp. 525-535.

Lashley, K. (1960) *The Neuropsychology of Lashley.* New York: McGraw-Hill.

Lettvin, J.Y., Maturana, H., Mccullogh, W. & Pitts, W. (1959) What the Frog's Eye Tells the Frog's Brain. In W. McCullogh, *Embodiments of Mind.* Second Printing (1989) Cambridge: MIT Press.

Levitt, L., Lewis, D. A., Yoshioka, T. & Lund, J. S. (1993) Topography of Pyramidal Neuron Intrinsic Connections in Macaque Monkey

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Prefrontal Cortex (Areas 9 and 46). *Journal of Comparative Neurology*, 338: pp. 360-376.

Llinas, R. R. & Ribary, U. (1992) Rostrocaudal Scan in Human Brain: A Global Characteristic of the 40-Hz Response During Sensory Input. In Basar, E. and Bullock, T. (Eds.) *Induced Rhythms in the Brain*. Boston: Birkhouser.

Llinas, R. R. & Ribary, U. (1994) Perception as an Oneiric-like State Modulated by the Senses. In: KOCH, C. E. & DAVIS, J. L. (Eds.) *Large-Scale Neuronal Theories of the Brain*. Cambridge: MIT Press.

McCloskey, D. I. (1977) Corollary Discharges: Motor Commands and Perception. In V.B. Brooks (Ed.) *Handbook of Physiology*, The Nervous System, Vol. 2. Bethesda: American Psysiological Society.

Melchitzky, D. S., Sesack, S. R., Pucak, M. L. & Lewis, D. A. (1998) Synaptic Targets of Pyramical Neurons Providing Intrinsic Horizontal Connections in Monkey Prefrontal Cortex. *Journal of Comparative Neurology* 390: pp. 211-224.

Mishkin, M. And Murray, E. A. (1994) Stimulus Recognition. *Current Opinion in Neurobiology* **4**, pp. 200-206.

Penfield, W. & Boldrey, E. (1937) Somatic Motor Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation. *Brain* **60**: pp. 389-443.

Phillips, W. A. (1997) Theories of Cortical Computation. In Rugg, M.A. (Ed.) *Cognitive Neuroscience*. Cambridge: MIT Press.

Phillips, W. A. & Singer. W. (1997) In Search of Common Foundations for Cortical Computation. *Behavioral and Brain Sciences*, 20: pp. 657-722.

Ringach, D. L., Hawken, M. J. & Shapley, R. (1997) Dynamics of Orientation Tuning in Macaque Primary Visual Cortex. *Nature* **387** (6630): pp. 281-284.

Ringach, D. L. (1998) Tuning of Orientation Detectors in Human Vision. *Vision Research* **38** (7): pp. 963-972.

Schneider, G. (1969) Two Visual Systems: brain mechanisms for localization and discrimination are dissociated by tectal and cortical lesions. *Science* 163, pp. 895-902.

Shepherd. G. M., Brayton, R. K., Miller, J. P., Segey. I., Rindsel, J. & Rall, W. (1985) Signal Enhancement in Distal Cortical Dendrites by Means of Interactions Between Active Dendritic Spines. *Proceedings of the National Academy of Sciences USA* **82**, pp. 2192-2195.

Shepherd, G. M. & Koch, C. (1990) Introduction to Synaptic Circuits. In G. M. Shepherd (Ed.) *The Synaptic Organization of the Brain*. Oxford: Oxford University Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Singer, W. (1990) Search for Coherence: a Basic Principle of Cortical Self-Organization. In: *Concepts in Neuroscience* **1**, pp. 1-26.

—. (1993) Synchronization of Cortical Activity and Its Putative Role in Information Processing and Learning. *Annu. Rev. Physiol.* 55, pp. 349-374.

Smith, E. E. & Jonides, J. (1997) Working Memory: A View from Neuroimaging. *Cognitive Psychology* **33**, pp. 5-42.

Sparks, D. L. & Groh, J. M. (1995) The Superior Colliculus: a Window for Viewing Issues in Integrative Neuroscience. In M. Gazzaniga (Ed.) *The Cognitive Neurosciences*. Cambridge: MIT Press.

Sperry, R. W. (1950) Neural Basis of the Spontaneous Optokinetic Response. *Journal of Comparative Physiology* **43**, pp. 482-489.

Sperry, R. W. & Miner, N. (1955) Pattern Perception Following Insertion of Mica Plates into the Visual Cortex. *Journal of Comparative and Physiological Psychology* **48**, pp. 463-469.

Stein, B. E. And Meredith, M. A. (1993) *The Merging of the Senses*. Cambridge: MIT Press.

Tanaka, K., Saito, H., Fukada, Y. & Moriya, M. (1990) Integration of Form, Texture, and Color Information in the Inferotemporal Cortex of the Macaque. In Iwai, E. and Mishkin, M. (Eds.) *Vision, Memory and the Temporal Lobe*. New York: Elsevier.

Tanaka, K. (1993) Neuronal Mechanisms of Object Recognition. *Science* **262**, pp.685-688.

Teuber, H. L. (1960) Perception. In: J. Field, H.W. Magoun, and V.E. Hall (Eds.) *Handbook of Physiology*, Sect. I, Vol. III. Washington: American Physiological Society, pp. 1595-1668.

Trehub, A. (1997) Sparse Coding of Faces in a Neuronal Model: Interpreting Cell Population Response in Object Recognition. In J. W. Donohoe and V. P. Dorsel (Eds.) *Neural-Network Models of Cognition*.

Ungerleider, L. G. & Haxby, J. V. (1994) 'What' and 'Where" in the Human Brain. *Current Opinion in Neurobiology* **4**, pp. 157-165.

White, E. L.(1989) *Cortical Circuits: Synaptic Organization of the Cerebral Cortex - Structure, Function and Theory*. Boston: Birkhauser.

Yang, C. R. and Seamans, J. K. (1996) Dopamine D1 Receptor Actions in Layers V-VI Rat Prefrontal Cortex Neurons In Vitro: Modulation of Dendritic-Somatic Signal Integration. *The Journal of Neuroscience*, 16 (5): pp. 1922-1935.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER TWELVE

THE MYTH OF NEUROCARTOGRAPHY

JOÃO DE FERNANDES TEIXEIRA

(TRANS. DIANA NEIVA)

There is a discipline that grew exponentially in the last decade of the 20th century: neuroscience. Its growing is, however, strange, for it fights against the destruction of certainties and ultimate truths that characterize post-modernism in general.

In this sense, if there is a theme today that can (and should) concern psychologists, it is the relation between psychology and neuroscience. This concern is legitimate and is due either to biopsychiatry advances, or to the advances towards a psychoneural reduction specially promoted by neuroimage techniques. We would be stepping into the golden age of neuroscience predicted by the so called eliminative materialists such as Paul Churchland who proclaimed the temporariness and the imminent end of psychology. From alchemy we would have passed to chemistry, and in the same way from psychology we would now be heading towards neuroscience.

The brain would have become, finally, the philosopher's stone, the key to all post-modern Science. Recently, with no surprise, I've read a text on the internet about the eminent neuroscientist V. S. Ramachandran (author of the best-seller "Ghost in the Brain") in which he claimed that neuroscience is the 21st century new philosophy, and thus it will soon occupy the place of philosophy, given that through the brain we could explain the nature of all our problems, even personal, existential and social ones – including in these the colonialism, which, as a native Indian, always concerned him. The colonizer's brain would be different from the colonized one, and maybe that would be the key to explain the triumph of colonialism through centuries.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the same direction, the philosopher and neuroscientist John Bickle proposed the reduction of traditional philosophical problems such as, for example, the one of the nature of knowledge, to brain structures. Neuroscience would be the end of philosophical problems, as we were progressively localizing the brain spots responsible for the skeptical doubts. We would also find the brain spot of moral (see the book "Hardwired Behavior" by Laurence Tancredi; and more recently "The Ethical Brain" by Michael Gazzaniga) beginning the neuroethics and, in a close future, the neurotheology, as we can localize the brain regions responsible for religious faith. The brain would be more than a philosopher's stone: it would be Aleph from Jorge Luís Borges, for it would be not only an object in the universe, it would be the universe itself.

But could neuroscience really dissolve the philosophical questions? What philosophical questions would be left after we unveiled brain mechanisms of perception, of knowledge, emotions and anguish? What is the distance between Angst (existential anguish) of Dasein and the relieve provided by a Lexotan pill? To the neuroscientist, anguish is just a brain's phenomenon; its causes don't matter in the so called "real world", and its solution depends exclusively upon an internal change of the organism. It doesn't matter if the economy is right, what matters is that it is said the economy is right. Along the same lines, it doesn't matter if the world is bearable, what does matter is that we feel and think it is bearable.

The presupposition of all these statements is based, however, in the project of brain mapping initiated in the 90's decade, with the invention of neuroimage, through CT scan and fMRI. Neuroscience would be reviving the project of neurocartography, something equivalent to reviving phrenology – a kind of electronic phrenology (initiated in the early 19th century, phrenology was the attempt to localize brain functions through cranial format, by palpation. So it was also called cranioscopy. The movie "The Enigma of Kasper Hauser", by Herzog, has some interesting phrenological scenes).

And yet is not it considered odd, neither causes any uneasiness reducing phenomena like moral behavior or religious faith – and even colonialism – to brain parts some would have and others not? Aren't there any methodological and maybe epistemological difficulties in neuroscience to execute this project? How could we clarify this which causes oddness and confusion? To evaluate this question more accurately it is necessary to go back a little in history or in pre-history of psychology and neuroscience.

In the late 18th century, German philosopher Immanuel Kant published a book (*Critique of Pure Reason*) maybe more important than the French

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

revolution which was happening. Way before the existence of psychology as a discipline, he alerted to the dangers of making a science of the mind. The pointed out problem was this: how could mind know itself? And could we know its ultimate nature, would the mind we're able to know be different from the mind itself? To know how the mind really is, it would be necessary to have a transcendent knowledge, a cosmic chair from where we could adopt a privileged point of view to know if subjective experience can or cannot be reduced to brain or neurophysiology, or even to know to what extent one thing is connected to the other. Our own mind doesn't give us that privileged point of view, which is cognitively inaccessible because the best we can do is to represent our mind using itself, which would then prevent a definite knowledge of the nature of the mind.

The mind we can know is the mind as it appears to us, filtered by our own cognitive apparatus. For example, to understand our mind as forming a unit is just a way of presentation of subjectivity to itself. On the other hand, we only have access to the way our mind functions (thought's contents) and not to the brain mechanisms underlying that functioning. These are never included in the thought. We are a black box to ourselves.

For we don't have any access to brain mechanisms, mind is shown to us as independent from these or independent from its underlying body. This is where it arises the inevitable illusion of immortality of the soul or the mind's persistence in time after the supporting body is destroyed. Immortality or the feeling of immortality of the mind would be, for Kant, an inevitable illusion.

Psychology as a science, in Kant's view, would be almost impossible. It would only remain the philosophical anthropology, something that in the limit would be approximate to the project of the psychology of peoples – this unread part from Wundt's work who tried to turn psychology into a science, almost 100 years after Kant. But this is the forgotten Wundt. The Wundt to whom we have access is the one read and interpreted by Tichener, so by the American interpretations of history of psychology which privileged, in Wundt's work, the recovery of introspectionism. This interpretation seems to have swept the kantian critic under the carpet, which, it seems, would have been deliberately ignored. Psychology would have freed itself from the inconvenient question that arises the suspicion that it could not be more than a building constructed over clay feet, fake foundations until today.

The fundamental question of Kant is raised again when we question, nowadays, if the brain is capable of knowing itself. This other way to formulate the skeptic kantian doubt should preoccupy neuroscientists and maybe force them to abandon their philosophically naive position. An

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

unforgivable naivety, despite the historically tense relationship between neuroscience and philosophy. Let us not forget, for example, that Hegel strongly attacked Gall, the father of phrenology, saying that "intelligence is not a bone". And Bertrand Russel, in a book published in the begging of the 20th century, "The Analysis of Mind", suggested that the representation of the brain is a product of the brain, which could, at most, be understood as a statement that psychology and neuroscience don't do anything more than to go in circles when they try to solve the problem of the mind-brain relation.

After all, what can the brain know about itself? Or also: can neuroscience help us to definitely relate subjective experience with neural representation? As we can see, it's not only about studying the brain as we study any other body's organ, and this is what gives specificity, but at the same time difficulty to neuroscience: it puts us at a peculiar cognitive situation, constituting itself in a discipline whose object is also its author or inventor.

The risk of walking in circles without noticing it was pointed out for the second time in the history of psychology by Freud. In his Metaphyschology he (who slowly started publishing since 1913) invites us to abandon the real *topica* of mental for an abstract *topica*. The abstract *topica* means, among other things, that the mind can only be "represented" or "imagined" and never "observed" by itself. The abstract *topica* is just a consequence of what Freud had already pointed out in his famous and always cited chapter VII of "The Interpretation of Dreams".

This chapter is contemporaneously criticized for having an uncouth neurophysiological model. But it is precisely there where the most intriguing aspect lies. Actually, the key to understand this chapter is at the passage where Freud tells us about ganglia and fantasies, where it can be observed that maybe to some extent we don't have a distinction between "ganglia of fantasy" and "fantasy of ganglia" (this pun was suggested by my student César Xavier). This indistinctness suggests a circularity where we wouldn't be correlating thoughts with brain areas, but just with other thoughts, these last ones representing brain areas. This is why Freud's psychic model can be fanciful or metaphoric not mattering its correspondence with a real brain. (Moreover, brain metaphors are always suspicious, as for example metaphors of spatial type which localize "high cognitive functions" at the top of the pyramid, and the "lower" at its base – but who can guarantee us that that is how the brain organized them?)

Freud's fanciful neuroscience is, however, totally permissible by its heuristic value, even if it is dissonant with empirical investigations. (This is not what occurs, for example, with *Totem and Tabu*, where the empirical

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

dissonance makes the text a manual of fantastic anthropology, as Levy-Strauss described in *Totemism*.) So it is about imagining the brain as a virtual machine that can execute certain functions – and the virtual machine doesn't need to be neurologically realistic. In this sense, movements such as Solm's neuropsychoanalysis are regressive in the history of psychology, for they ignore Freud's own texts where he already told neurocartography was just a myth.

***

But is this neurocartography possibility a consensus among contemporary neuroscientists? Or did they overcome past critics? The problems neuroscience faces to get this achievement are not only philosophical or principal objections, but also methodological objections, many of them raised by the American psychologist, William Uttall in his book *The New Phrenology*, published in 2001. He recognizes the merits of contemporary neuroscience for having turned into "cognitive" neuroscience through the observation of brain phenomena "in vivo" enabled by neuroimage. But, at the same time, he warns about some difficulties which seem to be overshadowed by the enthusiasm of the brain's decade.

The main difficulty consists in the impossibility of establishing a univocal taxonomy (a classification) for the mental. The difficulty is not the brain mapping, but the mapping of mind that this presupposes as a preliminary task. And the mapping of the mental – prerequisite to find its neural correlates – is based only in a cognitive version we have of our mental states; a version in mentalist vocabulary, using the rudimentary and imprecise theoretical instruments of our folk psychology.

This was a difficulty already felt by Gall himself in his phrenology. He distinguished twenty-seven abstract capacities such as individuality, benevolence, hope, self-esteem, etc. Cognitive neuroscience seems to have entered a similar adventure trying to build a map correlating mind and brain with new and highly sophisticated instruments, but taking as assumptions psychological concepts and entities derived from 19th century psychology and common sense. Trying to find neural correlates of such ethereal entities such as intelligence, consciousness, humility, desperation and other concepts formed by our common language and that impregnate psychological theories can become a task as ungrateful as to photograph the Tropic of Capricorn. Ironically, we would have to talk about a Damásio's error here – so we don't unfairly pay attention only the Descartes' error – this means, the fact that the contemporary neuroscience maintains the castesian theme. We would be reviving, this time in

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

neuroscience, the old sentence with which Wittgenstein attacked psychology by saying that, despite the experimental methods, it suffers from conceptual confusion.

Wittgenstein also saw reductionism with disdain, asking if there would be independent concepts of psychological states that conveys them. Is the proposition 12+12=24 true independently of the state of my neurons in the moment when I think about it, even when I, for example, have a fever? And therefore, are there independent concepts of brain states which convey them? Are true propositions correspondent to the activation of a brain's region, and the false propositions to another? Do truth and falsehood depend on the activation of a brain region and not the other? Or are they independent concepts of psychological and brain states – something that can't have a neural representation precisely for the fact that "true" and "false" predicates don't correspond to anything in the world?

Even if the reductionist could prove there were regions in the brain responsible for making certain propositions into true or false ones, how would the brain represent a sentence which is always true as the famous Descartes' statement "I think, therefore I am"? This is a proposition whose denial is also true. In this case, which brain areas should flash, the ones corresponding to true or false propositions? Or is there a special region in the brain responsible for the representation of peculiar propositions and paradoxes?

The same maybe applies to the predicate "being conscious", which we want to attribute to some of our mental states. Philosophy of mind, for some time, started looking for the so called "neural correlates of consciousness". It would be a special set of neurons, or maybe special some characteristic about them, responsible for turning some mental states in "conscious states". And many hypotheses in this direction were formulated by cognitive neuroscience in the last years. All of them bump into the necessity to previously conceptualize what would be understood as consciousness before starting to hunt the neural correlates down. The factory of theories of consciousness runs at full speed with great annual productivity. Soon, formulating a new theory of consciousness will be the task for a monograph to graduation in USA.

This search for neural correlates of consciousness which occurs today resembles the situation described in a short story by Argentine Julio Cortazar, in which a family starts to dismantle its apartment in search of a strand of hair with a knot. After they dismantle the whole apartment, and without finding it, they decide to work and save money for several years to be able to buy the apartment of the neighbor downstairs. And when they do this, they start to eagerly dismantle it to find the strand of hair with a

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

knot. Finally, they find it. But then a terrifying doubt rises: is this the strand of hair we were looking for? Or is it another one?

The considerations above lead us to the conclusion that the history of a person's mental life cannot be mapped with precision in the history of events in the body of that person – a conclusion that disturbs the brain mapping project. If we can't map truth values and correlates of consciousness neither it would be precise to say, for example, that a person knows or believes this or that. In other words, the terms that describe private events are inaccurate, for there are no connections in the nervous system that conduct the sensory nerves to the right places – a fact that is pointed out by Skinner in various passages of his work. (Moreover, it is the recognition of this fact that evidences psychology's necessity to contain, besides other topics and objects, a behavior science – a science with its own sphere in which the psyche is approached in an externalist and relational perspective.)

We stand before a critic to the Cartesian psycho-physic myth, the myth that there is a two-way correspondence between mental and brain events. Descartes was the first to affirm this myth, and the first one to sustain that thought occurs in the brain – we find these assertions in passages of his late work, *Principles of Philosophy*.

Besides, the correlation between brain areas and cognitive functions requires the assumption of the possibility of a methodological divisibility of the mental. So, is the possibility of division of the mental, even if it is just methodologically assumed, a sustainable premise? Or, in other words, can we assume the modularity of mind – even in a version softer than the one sustained by contemporary cognitive scientists such as Fodor and Pinker? Which criteria should be established to relate mental modules with brain modules? And which criteria should we follow to establish brain modules? (Because it's supposed that they are, to adopt these gentleman's vocabulary, “encapsulated”, and only a physical thing could be encapsulated.)

Are there more problems? Yes, and now I want to give an opportunity to hermeneuts and all those who believe in the own reality of the meaning. Is neuroimage useful to them? Or, in other words, is it possible to detect and “photograph” the sense? We should cite the experiments of Wise and Chollett performed in 1991 which showed that the same areas flash when we present a sequence of words or of letters that don't make any sense to the subject. An experiment cognitive neuroscientists would rather let be unnoticed, for it shows that the sense is not detectable, and even lesser reducible to a “thing”.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Is it about attacking neuroscience? Neuroimage? No. It's about attacking its totalitarian and reductionist pretensions. It is necessary to rebalance the scale, rethinking its interdisciplinarity in the composition of a general science of mind. This is a task I tried to engage in my book *Philosophy of Mind: neuroscience, cognition and behavior*, rethinking radical behaviorism under the light of cognitive neuroscience. It is necessary to criticize science, so it can make progress, so we don't incur on what happened to psychoanalysis, which never advanced any more for fearing to criticize its master.

Psychology will always defy the ideal of a unified science, since the construction of psychological explanations will always require the re-combination of distinct theoretical perspectives such as the case I focused on, between cognitive neuroscience and radical behaviorism. If the peripheral explanation of radical behaviorism doesn't exhaust the diversity of cognitive and behavioral phenomena, either neuroscience could do it on its own. The richness of psychology lies here and not in its pre-paradigmatic poverty, as some suppose. The unifying ideal should have already been banned from post-modern science long ago. That ideal is inspired by physics, which in the beginning of the 20th century was considered the model of science – of the science that vainly searches for a unified theory of the universe. An ideal whose ultimate inspiration is the monotheism that underlies the Christian thought.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER THIRTEEN

SCIENTIFIC DREAMS AND FOCUS FICTIONS ON CONSCIOUSNESS

JUDITE ZAMITH-CRUZ
AND ANDRÉ ZAMITH CARDOSO

Life experience translates into psychological reality through memories, imaginations, visions, smells, touches as sensory perceptions, cognitive and affective processes. In a person's focusing ability, someone does not live the consciousness of someone else. With 3D brain images, functional magnetic resonance imaging (fMRI) has been used to correlate and reconstruct the active parts of the brain with what is being observed by the subject (Naselaris et al., 2009). In human beings as a whole, the complex and the superior are included in the frame of *extended consciousness*, developed by Damásio (1999). If consciousness plays a key role in allowing us to bring information together in novel ways, researchers suggest that consciousness serves other functions too. Edelman et al. (2000), Nir et al. (2010) or Graziano (2013) have expanded the knowledge about the dynamic and correlational mental processes. Nili et al. (2015) questioned our imagination: "How does an information-processing machine produce subjective awareness?" The understanding of consciousness is discussed in five paradigm shifts. Here, a literary and phenomenological view on consciousness is examined, as well as a psychobiological, neurological, emotional, social-neurological perspective. The thoughts on consciousness that stem from science-fiction films and literature are analysed in the context of building an artificial intelligence (AI) machine.

Introduction

The mind interprets reality through generating mental representations. In this sense, representation has the sense of "product of mental activities

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

exerted on the real" (Tiberghien et al., 2002), that are neither copies nor intrinsic characteristics of the reality.

Some years ago, Carter (1998) described the cognitive thinking process as a *working place* of consciousness. There, the metaphor of the processing features of consciousness appears as a symbiosis of cortical areas in the frontal lobes with "evaluating tasks", while subcortical areas feed these areas as "underground production chains". This underground subconscious process is the largest producer of emotions, which have an enormous effect on the unconscious experience, densely connected to the grey matter. Consequently, it has not been possible to *see* consciousness by brain imaging techniques; furthermore, we cannot delineate areas of memory, images or thoughts.

There are numerous theories to explain consciousness (Tiberghien et al., 2002): activation (exceeding a certain threshold for a mental process to be conscious), novelty (the requirement of new information to be brought to consciousness), *the tip of iceberg* (the imposition of a set of emergent exchange of unaware experiences to become aware), and *the theatre* (the needed place to collect the information in order to make a person aware). Baars (2002) and Bennett et al. (2007) defend the idea of "information" (novelty), when they think of consciousness as a "series of input and output that form a chain where information moves on". As the stream of consciousness is erratic and fragmentary, Dennett is outside of the theatre metaphor. Rosenthal (2005) suggests a new "quality-space theory", part of the global-workspace theories. Graziano (2013) suggests the question is about social neuroscience and the attention process: "How does an information-processing machine produce subjective awareness?" Does social perception give a person the feature of awareness that can be attributed to someone else?

What is the interest in consciousness and in the knowledge of existence? Consciousness is present in our perception of the world and in the representation of other people. It emerges from the knowledge of feelings and the cognition process. These combined processes are present in self-consciousness, and it is connected with the intuitive sense of experiencing its meaning. This happens either when we connect perceived memories and imaginations and when we are offered the sense of living.

The prefrontal lobes (stress control, dithering and planning tasks) and cortex are essential in conscious acts (Crick and Koch, 1998): it was found that the type of neural mechanisms that underlie the organization of visual perception, and the conscious perception of emotions and focused attention are related to prefrontal cortex activity.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The evolutionary theories suggest that if human ancestors and other hominids (extinct up to one million years ago) evolved, there should be neuroscientific evidence about the adaptive value of consciousness (Rossano, 2003; Striedter, 2005; and Rakic, 2010). It was in the evolution from hominids to human that the frontal lobes increased to approximately 28% of cortical area of the brain (Carter, 1998), and consciousness widened in the *core consciousness*. The brain expanded to incorporate something that may well be only human – the difference between to be (passive *automata*) and to act, the appearance of self-determination (to follow needs, demands and desires). The extinct animals possessed brain regions focused on the neocortex; however these were located in different places as compared to current mammals (Mlodinow, 2012).

Nowadays, images of the brain not only show the mechanisms of aggression and the systems involved in perceptive dysfunctions (Logothetis, 2008), but also mood, guilt and self-esteem (Wagner et al., 2011). The brain forms images of the body and external objects, creating a second order representation. This is not an abstraction with regard to self-consciousness. In brain mapping, the representation of the self-body is activated in the hypothalamus and cingulate cortex (Damásio et al., 1996). Several brain regions are assigned to the conscience task of making the body aware. Simultaneously, consciousness is also controlling awareness of the rational process under way. An area that appears to control consciousness is the ventromedial cortex (Drevets et al., 1997), also associated with depressive states. A disruption in activity in this area makes us perceive a lack of meaning in life or fall into manic states. This is a productive emotional centre, with several circuits of cognition and emotion. Those regions are extremely dense, and unite the conscious and unconscious mind. Unconscious processes gather different "results", which are unknown, however consciousness recreates these "products" (Tiberghien et al., 2002). *Freewill* depends on the selective function of consciousness, we can select from random elements which one is "interesting" to think about. Freewill is associated with the orbitofrontal cortex (conduit adapter to fluctuations in social and emotional context), below the ventromedial cortex. These were linked to reward and to hedonic experiences by fMRI (Kringelbach, 2005). In these cases the function of the brain is to recreate, construct and modify the experience, rather than receiving it. In harmony with the shared experience, when someone is asked where they feel their focus is, they usually point to the region just above the nose (Carter, 1998). Despite not being aware of this or other brain regions, they still point to the prefrontal cortex region.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The state of consciousness that occurs during the *waking life* is also called "waking consciousness". At this point, a number of questions arise: How to escape the approach of that "normal" consciousness, a moral entity or public conscience? How is it possible that most of our illusions occur without unveiling different levels of consciousness?

In this paper, we intend to discuss the interpretations of the conscious mind, and some apparent lack of consciousness: attention deficit hyperactivity disorder, fainting, hysterical paralysis, coma and *blindsight* (Celesia, 2010; Weiskrang, 1986). Additionally, the state of a professional athlete is examined (e.g. tennis, baseball, badminton, cricket, hockey) when they get to instinctively react, for example, to a high-speed approaching ball. The case of a person with altered states of consciousness (ASC) such as sleep deprivation, hallucinogens and mental disorders, usually linked to a deficit in brain states, is discussed in conjunction with a possible brain injury or chemical/biological molecular imbalances. In neuroscience, determination and self-consciousness were studied by interventions conducted in war veterans and other patients (Doidge, 2007; Carey, 2008), who suffered serious injuries in the pre-frontal lobes, with notable personality changes.

XXI Century Approaches: Theories and Beliefs

In neuroscience and cognitive psychology, there were three questions of initial focus: "How to connect consciously the large number of perceptive and psychical processes in a single coherent whole?" (Crick, 1994; Llinas et. al., 1994; Singer, 1998; von der Marsburg, 2002); "Is consciousness determinant for self-awareness (Keenan et al., 2001), and does it assist emotional control (Silvia, 2002)?" Other influential theory (Baars, 2002) suggests that consciousness brings us "information" by novel ways. In an intuitive view consciousness is only an "internal light bulb illuminating the mind".

Recently experiments were carried out on the introspection on consciousness (Natsoulas, 2001; Varela & Shear, 1999). These resulted in a division between an inherently *easy*, and a *hard* kind of "problem". In order to give meaning to subjective experience, psychiatry uses new labels originally introduced to describe other syndromes (e.g. schizophrenia, among others). These medical terms are currently associated with ASC. The psychology of consciousness uses most of the same labels, such as in psychiatry.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

As the concepts are based in beliefs (epistemic attitude about uncertainties), the theories are partly based in beliefs, subject to historic and linguistic turnarounds. Beliefs are metaphorical artefacts, which we need in order to predict the behaviour. They can be found both in superstitions and prejudices. Functionalism unfolds on a belief, for cognitive processes located in the social structure we live in. The beliefs system may (or may not) be contradictory, absurd, or contradicted by scientific evidence. Such an example of belief is when each of us feels more comfortable believing that we are in control of our own lives. According to the attribution theory (Lerner, 1980), humans are led to believe in a *Just World*, where it is possible to predict and guide our pathway. Beliefs are rooted and are extremely difficult to change, as it is difficult to change values, conceptions of self, the notion of reality or power.

Some people do not agree with most of other's beliefs, knowledge, and feelings (in greater extent, people with autistic type of characteristics). Identity and inter-subjectivity in schizophrenia and autism have been studied in clinical psychology for many years.

One thought experiment about autism suggests that a "shared activation mechanism" is needed, at a motor level and at an intentional level, beyond the "theory of mind" (TM) associated with beliefs that fail in autism. In that perspective, a theory is not only a cognitive, intentional and logical process (Tiberghien et al., 2002). Several research groups have suggested that the TM can be observed in the unfocused look (Premack & Woodruff, 1978; Leslie, 1991; Frith, 1989; Lillard & Skibbe, 2004; Welborn & Lieberman, 2015; Baron-Cohen, 1995, 1997, 2011). *Mindblindness* and the absent *mind reading* are characteristic of autism (Baron-Cohen, 1995). Disrupted communication and lack of empathy are used to characterize autism. The subjects did not have "other minds in the brain" (Fletcher, et al., 1995). Some people cannot "get themselves in someone else's head" (TM) and do not believe intuitively that others have a different vision of the world from their own.

Frith and Happé conducted a study with positron emission tomography (PET scans) where they compared people with and without *Asperger* features (Baron-Cohen, 1997). The subjects were asked to infer the other person's mental state during the PET scans. The control group had an active median left prefrontal cortex. For *Asperger syndrome* subject group, a region located immediately below the one that became active for the control group was active instead. With functionalist registration, people who avoid eye contact (Baron-Cohen, 1997) use modes of "processing of information" that are often distorted, such as save and retrieve, combine

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and remember, for cognitive operations, generated by simple images, words or geometric figures stimuli.

As functionalism is a top-down view of brain functions, these researchers break down the functions into singular elements. Alternatively, the bottom-up view explains neuronal functions from the simple to the complex (from molecules, cells, brain, individual up to social networks). Therefore, perception of a functional mind features these two interpretations.

We can refer to folk psychology as a bottom-up theory. Folk psychology is an expression used to refer to behavioural reasons (an immediate human resource), as well as reasons for abstract concepts (beliefs and desires). Since those assumptions "originate" states and mental events, they are the "causes" of these behaviours. Folk psychology is ascending – bottom-up (Bruner, 1986, 1990, 1991). It starts from a few stories situated in a timeframe, and finishes in epochal dreams and blurred visions of consciousness. Narratives of knowledge and existence cannot be tested as hypotheses because people are not very predictable, the world is not fully knowable, and phenomena to be experienced in the world appear to be irreplaceable when these appear to be in controllable situations (Zamith-Cruz, 1996).

The Enlightened Land of *Psychologies*

The acts of reading, speaking, memorizing, thinking or performing an automated task are human attention activators, such as driving a car.

One of the most prolific researchers on attention and consciousness was Francis Crick (1916-2004). He surprised a psychologist with the manifest view of consciousness by a simple mental exercise (Kosslyn & Rosenberg, 2004): "Hold both hands in front of you, but with one closer to you. Now look at the front one; now look at the back one. See how the front seems different when you are focusing on the back one?" The psychologist did not like to hear that consciousness was merely an attention aspect of the brain, to what Crick have responded: "consciousness is enriched by attention, but attention is not the necessary awareness". What Crick and Koch (1985, 1998) exposed were the crucial areas of the frontal lobes to consciousness. According to them, consciousness does not arise from regions of activity that register perceptive information: the primary visual cortex (V1) and the primary auditory cortex (SMA V1). Grazziano (2013) pursues an "attention schema theory", thinking that the specialized machinery of the brain calculates the form of consciousness and gives it to others in a social context. Nowadays, Koch (2009, 2012) suggested the fundamental property of consciousness to be like mass-energy and

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

electrical charge expressed through local concentrations of "integrated information".

Consciousness can also be observed in mundane human situations. An intuitive situation is attention deficit, lack of concentration, hyperactivity (physical restlessness), uncontrollable impulses and failed integration of stimuli or inputs. However, usually the hyperactive child is a "difficult" one. They have irregular activity in the prefrontal cortex, the anterior cingulate gyrus (i.e. attention focus in the stimulus) and/or the upper auditory cortex (i.e. integrates stimuli from different sources). In attention deficit hyperactivity disorder (ADHD) the brain may not be fully activated. Therefore, while sub-cortical regions are fully functional, the prefrontal brain areas fail to work synchronously (Cowen et al., 2012). In this case, the dominant protagonists of consciousness fail in *sustained attention*.

It is common sense that if someone faints and is not consciously present, the person stays physically immobile; therefore the person loses their consciousness. These regions do not appear activated in PET scans, while the frontal lobes appear activated, in the case of a leg movement. So one can see that having inactive limbs can occur despite the fact that their corresponding brain connections of consciousness are still intact.

The full sense of the "phenomenological flow" sets up another illusion. When someone is in the state of vegetative coma, the person can give us the idea of being aware. When a hand is stung, the spinal cord and the thalamus immediately trigger the fixation of the eyes of someone else by a rapid gesture. Yet it is an automatic reflexive activity. A tennis player can act instinctively, unlike a blind person, he possesses brain activity on the occipital lobes; nonetheless it is not an automatic reaction only. A blind person can move their face rapidly to the attention direction. Thus by extension of meaning given to their conscience, the blind person will also come to interact with their attention seeker (Pegna et al., 2005). In imminent danger, a person even uses a *blindsight* (Celesia, 2010). It is possible that the study of consciousness may unveil more insight in the field of dementia (e.g. Parkinson's and Alzheimer's diseases) and strokes.

The Literary and Phenomenological Consciousness

In the phenomenological domain, Romanyshyn introduced a metaphorical aspect of experience. This was part of the extension of meaning, and similar to the reflection on experience by the classical philosophers Erasmus of Rotterdam (1469-1536) and Michel de Montaigne (1553-1592).

First, Romanyshyn argued that "from fidelity to the psychological experience, in its own terms, it takes us beyond the alternatives of facts

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and of ideas of things and thought, empirical and mental reality; it takes us to a metaphorical reality" (Romanyshyn ,1982; cit. by Becker, 1992). On the one hand, it is thought that experience comprises a kind of awareness. On the other hand, it is in the fictional realm that new theories have emerged beyond the ordinary cognition problems (the *easy problem* of consciousness).

The writer José Rodrigues Miguéis (1959) has fallen at the door of "ethical realism", a new strain of consciousness in the subjective and common sense "reality". He named hunger as the most urgent primary need (Maslow, 1954). It would be the resentment of an internal physical motivation (hunger), when he wrote about his character: "Suddenly I felt the guts rolled me up in hunger. (...) and sometimes there is nothing like such a simple desire to reactivate a man to his consciousness and confidence" (Miguéis, 1959). In Lev Vigotsky's original works of 1938, he mentions that there is a paradoxical demand of consciousness, in the case of hunger, it would cause social and political awareness.. He framed it in an internal model that mirrored his own condition, by saying that "the word is for the conscience as the small world is to the big world".

The First Scientific Turning Point: Functional Biopsychology

A function is generally opposed to a phenomenon. Alternatively to the phenomenological mind, the functionalist paradigm of experience is focused on the "products" or "results" of thought and it is not directed to the way of how we think during the (erratic and fragmentary) flux of thought. Instead the sensory activity is replaced by what happens in the brain. Could it be that activation of a neuronal module causes a *product* that is a private or subjective state of consciousness? Are the activation of action potentials in the thalamus and in the sensory-somatic cortex the main cause of the conscious perception?

A difference in perspective suggests that the structure of neural activity does not represent the stimulus. It is the person's meaning that is based on experience. The quality of *my* or *your* conscious experience has variations according to the cognitive processes involved (the various components of consciousness). Either it is the attention or the reflection on the qualities of things (*phenomenological consciousness*).

After the sixties, the cognitive paradigm of information processing, consciousness, is seen as a "function" due to the fact that we access it and we are effective in its transmission. Furnham (2008) outlined that we are aware of how information processing is consciousness. It is relevant to say

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

that in the functionalistic ideology, machines can have conscious. A fuller understanding of what it represents – what does it mean to be a person? – requires first understanding that consciousness must be considered in its environment. The main question is: why does an emotional experience precede consciousness? A revolutionary change leaded by Philiph Bard and Walter Cannon (1929) suggested that the hypothalamus is a key part of the emotional brain.

The major finding was that brain structure is different from a homogeneous black box (e.g. behaviouristic position). Canadian psychologist and biologist Roger W. Sperry (1968 and 1974) made the first inter-hemispheric surgical separation, cutting the corpus callosum, showing how the two hemispheres seemed to work independently. This procedure was first shown in small mammals and in severe epileptic humans (split-brain). Along with Ronald Myers, Sperry showed that a cat could learn a task (pulling a lever to receive food), while having a half of the brain activity ignored by other half of the brain.

How do two split minds cohabitate in the brain? A well-studied case is the one of a butcher with absence of inter-hemispheric communication. Usually, if a person lost the command of his left hemisphere (Springer & Deutsch, 1993), that person loses the control of one of the hands. That suggests that the brain architecture has differences between hemispheres. Epilepsy patients can have inter-hemispheric separation surgery. Fewer are those who have suffered strokes with brain lesions in only one or both hemispheres in a specific supplementary motor area (SMA) and the corpus callosum. In this case, the person perceives what is called a “conflict between hands” (Gazzaniga, 1967) or “a strange hand”, a disease known by patient M.P. (Parkin, 1996). In this case, M.P. was increasingly able to perform daily activities. Despite the fact that her left hand had "helped her" making a tortilla, it prevents her from fulfilling daily tasks on several occasions. This phenomenon is not exclusive of schizophrenia, here called "split personality”.

The Second Turning Point: Integrated Neuroscience

The phenomena of double consciousness can also be thought of following Damásio’s terms (1999): *consciousness* (the phenomenological flow, the temporal course of consciousness and its contents); and *consciousness of the self* (and the sensation of self), thus the sense of phenomenology to an entity (the *self* who feels *being*). We think and feel ourselves thinking (metacognition).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

In early 1960 cognitive neuroscience adhered to the concept of consciousness as a continuum of bodily manifestations, which are heterogeneous phenomena at a perceptive and reflective level, and, as mentioned, the way each body experiences life (subjectivity).

In regards to self-consciousness, Damásio interprets it as a narrative self, within the character caught up in *the stage*, when someone adopts different identities (Mlodinow, 2012). It was by leaving the identity of the self that Damásio (1999) suggested another reality on the cortical level and on the subcortical level. The theory expands consciousness in terms of intentionality, as our experiences have meaning. The representation of an experience and the wide range of experiences we have are connected. Damásio (1999) emphasized the current situation in *storytelling*, a function of the brain that captures intentionality, i.e. the fact that the psychic contents "relate to... (…something external to the mind)". This does not imply examining the "existence" in neurosciences. This is in the sense that consciousness is the knowledge we have of our existence. This does not imply that we should not study an *extended consciousness*, which Damásio mentioned by outlining *the human*, *the complex* and *the* "*superior*".

The "revelation of consciousness" was also presented by Damásio (1999). Here, biological mechanisms regulated by the body were introduced (associated to emotions). The aspect of self-regulation was also envisioned due to the relationship with the environment. Consequently, civilization is not an *extended consciousness,* which occurs in the minds provided with *core consciousness*. The interaction with social media is attributed to the collective minds of emotion, memory, language and intelligence. Secondly, these aspects suggested that consciousness was an *extended consciousness* and the concept of *self*. The new metaphor for the brain had a captivating *audience* of representations of *themselves*, as it included the body, constantly supporting its storytelling. Since then, consciousness includes the sense of *self*, coupled to the act of *knowing*. These theories identify a *thinking being* ability to *perceive* what is knowledgeable, and the knowledge of him or herself.

The Third Turning Point: Unconscious Emotions

> Nobody enjoys failing to explain their own rationales or intentions (Gazzaniga, 1992).

Since the eighties, the models of information processing integrated correlated emotions as cognition. Frijda (2007) and other scientists formulated that there is no true distinction between emotion (with certain general rules) and reason. Through the cognitive theory of emotions,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

without simple wishes or cognitive states, emotions would be over determined by deliberate intentions – “the readiness of states in action”. So, emotions are not "passions" opposite to "actions". They are accompanied by subjective feelings, giving clues to others about our thoughts and desires. Are there emotions disconnected from the bases of our reactions, in conjunction with an overall situation? Emotions are not always conscious intentions.

Nowadays, emotions constitute a primary system of meaning, and intentions are mixed beliefs and desires. The evolution of emotions involves that they are not "derangements" of human “positive” development, communication and interaction.

Joseph LeDoux has extensively studied fear conditioning, in the processes involved by two different brain mechanisms: a reflex system and a dependent system of thought and interpretation. In 1996, he showed emotional outbursts in animals, such as "false anger" because, as Walter Cannon would have said, they do not have a "conscious feeling of anger" (Cannon 1929, cit. by LeDoux, 1996). LeDoux (1996, 2012) changed the explanation on emotions, in agreement with the theories of James-Lange (we react to a stimulus and then we feel an emotion, only after the reaction), and Cannon-Bard (we have separate bodily reactions and emotional reactions). Thus, he was the first one to concisely explain emotions, from the perspective of William James, one hundred years after William James (1884). He devised a new circuit of fear named as *shortcut of LeDoux* ("we run away, and then we get afraid"), in which the passage of the input is in the order of thousands of seconds faster than the long circuit – "we have fear, and then we run away" (Fig. 13-1). His revolutionary breakthrough moment was based on works of Gazzaniga (1967, 1998); he extended on the knowledge of the "conflict between hands". He described the split-brain condition. While the patient beats his wife with one hand, he would protect her with the other hand. After all, we can carry a *mute prisoner* in our head (the right hemisphere), with a *distinct personality* from the everyday personality we have or we want to be (LeDoux, et al., 1977). The suggestion was that the left hemisphere controls most of the aspects of language processing, reasoning, and storytelling. The right hemisphere works in absence of interpretation of the world, and justified beliefs. The left hemisphere for its part gives us our “motivated reasoning”, an extremely positive self-image, as we believe in our goodness, competence or control (Mlodinow, 2012).

As manifestations of split-brain were known with better technical resources, Gazzaniga used the absence of inter-hemispheric communication to explore functions of the sensory and motor cortex on the control of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

emotional and unconscious behaviours (LeDoux, 1996). Experiments on the full cerebral cortex of cats which had their cortex removed resulted in indicators of emotional arousal, once a cat was provoked: they cowered, bent the back, meowed, retracted the ears, bristled their fur, showed their claws and their teeth, they bit objects when presented by hand, among other reactions (Kaada, 1960, 1967; cit. by J. LeDoux, 1996). Under these circumstances, decorticated cats activated the autonomous nervous system: their fur bristled, their pupils dilated, their blood pressure and heart rate increased. In conclusion, their behaviour was changed. They had unregulated emotional reactions and they were not able to control their animal rage. Therefore, it was impossible to continue accepting the rational that cortical regions controlled these emotional reactions (Head, 1921, cit. by J. LeDoux, 1996). Once these areas where removed, there was bilateral coordination loss.

The classical fear conditioning was based on the amygdala (detection and evaluation of the affective perceived content, emotionally modulating memory), a small structure on the frontal interior area of the temporal lobes. Thus, the emotional theory of Joseph LeDoux is an alternative to the dominant cognitive theories of emotion, with pioneers such as Richard Lazarus (stress research), Stanley Schachter (eating behaviour and cognition research) and Magda Arnold (relationships in personality and emotions research).

From the sixties until the end of the twentieth century, we believed that emotions emerged once a person interpreted a situation as a whole. The current idea was that the state of the body is affected by its surroundings; even if an emotion is "weak" and transitory, one would experience the intrusion of a thought (e.g. looking out of a window, meditating, one or two seconds). So, it was thought that there is a *control precedence* of an emotional occurrence (Frijda, 2007). However, we have different cognitive systems for different emotions. One cognitive system is acting as a reflex system (regardless of thought and interpretation), while the other system is functioning as dependent of thought and interpretation. On the one hand, a person has reactions acting upon the brain and body, while on the other the reactions act upon memories and interpretations, both systems involving emotion (Fig. 13-1).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

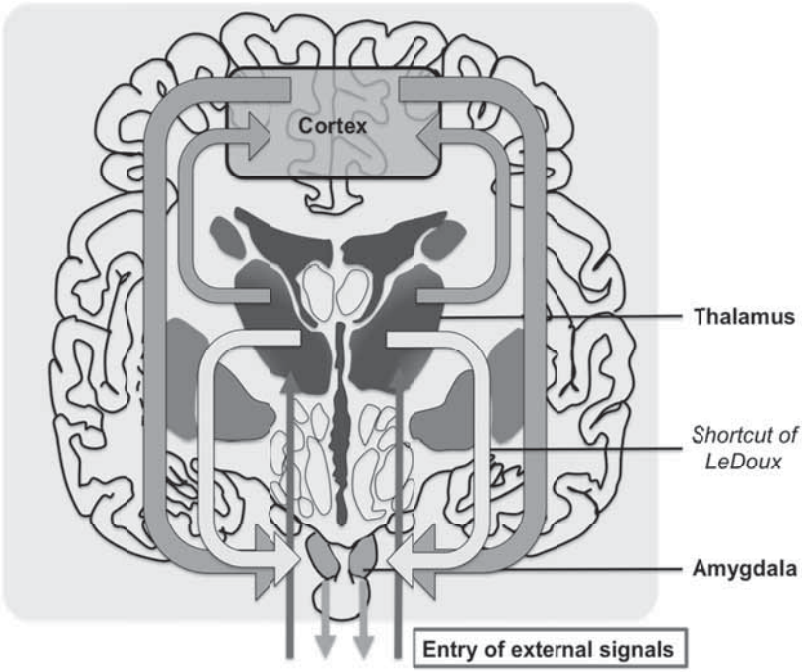

Fig. 13-1. Scheme of the flux of emotional information in the brain. The entry of external signals passes through the thalamus and can follow two paths: one through the amygdala or another through the cortex (conscious mind). The *shortcut of LeDoux* represents the shorter path through the amygdala, thus emotional unconscious reactions are faster than conscious ones.

When a baby reacts with an expressive-motor form (sensorimotor) to the environment that evokes their biologically adapted "answers", the baby already evaluates with the amygdala what is "good" or "bad" for themselves. Therefore, the fear connected to amygdala activation requires no cognitive interpretation. But complex guilt lies in cognitive interpretation and memories of previous events. By that (r)evolutionary approach (Amorapanth et al., 2000; Davidson & Begley, 2012), emotions arise at some occasions as brain-body interconnected reactions, while with some events they come up as part of conscious memories and interpretations of a given situation. The idea that emotions come from a deeper order than the frontal cortex, because there are more connections expanding from amygdala (unconscious emotional generator) to the cortex (medial temporal lobe, orbitofrontal cortex and the frontal lobe) than in the reverse

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

direction (from the cortex to the "limbic system") was innovative. Emotional control is imposed. The conscious mind is not at the center of emotion operations.

Biological or psychological emotional states are tested when we are unable to make cognitive decisions, and lack the impulse control due to frontal lesions (see Phineas Gage). Despite the large number of emotional connections to the frontal lobes, this does not mean that parts of the frontal lobes are not involved in motor control.

The Fourth Turning Point: The *New Unconscious*

In communication, we express different "types" of coherence. If communication gets across to the listener, it is legitimate, and we feel confident on our own identity of the self. The study of social self, conduct and relations exists within human civilization, a field nowadays called *social neuroscience* (which implies a structure within a previous frame).

Neuroscientists of the last four decades introduced the concept of "superior" level of the mind, including how automatic mental processes change us (Bargh, 2007). It suggests that the development of these working processes evolved for a better human adaptation (Wilson, 2002). These findings clearly show that animals not only have an instinctive behaviour, but also act beyond most common instinctive actions (Kolb & Whishaw, 2004; Mlodinow, 2012). Therefore human and other species share the beyond instinctive behaviour and neocortical tissues.

Recently, emotional unconscious terminology has changed, and it has been categorized in different dimensions: the implicit, the tacit, or the hidden mind. The work of Sigmund Freud (1856-1939) has been reassessed (Solms, 2004), with theoretical improvements, which resulted from the increasing molecular, cellular, neural and psychological research: such as in neuropsychoanalysis (Berlin, 2011) and *social* neuroscience (Galbis-Reig, 2004; Wilson, 2002).

Nowadays, most of the mental processes can be labeled as two types: conscious and unconscious (Kihlstrom et al., 1992). Unconscious processes led to the study of some mental processes such as visual perception (Barbur et al., 1993), "inattention blindness" (Levin & Simons, 1997), and false memory (Loftus, 1974, 2005; Levin & Simons, 1997; Simons & Levin 1989).

The birth of social neuroscience can be dated to a meeting in 2001 (Ochsner & Lieberman, 2001). Their research was not only advanced by the increase in novel imaging techniques (Naselaris et al., 2009), but also by the large amount of psycho-social studies included. In addition to the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

inherent subjectivity of who has to react to images or sounds (i.e. someone who is asked to think of *A* or *B*), in the context of brain research, the sophistication of human-like reaction is already well replicated by machines. Naselaris et al. (2009) used fMRI to monitor the flow of information in the brain when a person is asked to think of a place *A*. Subsequently, a computer is programmed to recreate with extreme precision (for a computer reconstruction) what the thought of the "real" image *A* represents.

First of all, these results raise the question of how the unconscious matrix is realigned with the emerging neuropsychological and social knowledge. Notably a third of the brain size is devoted to vision (sensory) and consciousness-associated systems (Mlodinow, 2012). In the unconscious vision, the temporal lobe becomes important to fill the gaps in our vision (as complementary to vision by the occipital lobe).

Secondly, unconscious processes are the result of operations performed as "logical machines" and its activity reflects the reflex domain (Baars, 1988). By that cognitive awareness perspective, the unconscious processes of consciousness perform very specific tasks fast, sometimes make little mistakes, and do not suffer interference from other processes, dealing with large amounts of data, operating in parallel, for specific and limited areas.

Thirdly, the intricate relation between the conscious and the unconscious suggests that we do not have the expectation that we control what we dream or what happens or goes around us (Hassin, et. al., 2005), simply because we do not have only a conscious mind (Mlodinow, 2012).

Our ignorance about the *hard* philosophical problem should remain, but it will only remain about "raw" experiences (*lived*), with fading traces at the level of immediate memory. One reason for this is that the experience we have is recreated by ambiguous theories of consciousness.

Finally, brain injuries and chemistry improve our understanding of consciousness, therefore we need to continue exploring more about it, also due to the frontal cortex being involved in dementia. Other conjectures come from computational developments, as we will see next (Nili et al., 2015).

The Fifth Turning Point: the Step to Conscious Machines

Human-machine interfaces (HMI) are welcome in society (Negroponte, 1989). The brain is not a silicon-based technological material, nevertheless it has its self-regeneration capacity and its sensory component in common with technology, in other words, to be a "mission control center"

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

(Eagleman, 2011). However, a silicon-based material acts as a gender-neutral "social actor" (Nass et al., 1994, 1997).

In an age of technological changes, machines put an end to behaviorism, the so-called "cognitive revolution" (Bruner, 1986; Gardner, 1986). Since the eighties, with *connectionism*, cognitive psychology has been associated to neurosciences, building the ideas of "executive routines/programs" and "mental models" (Johnson-Laird, 1983), with parallel processing and interdependent operations such as in computers. These methods of brain analysis came to replace the sequential, linear or serial operations (e.g. *associationism*). Following this, the structure of the nervous system would become decentralized, to be either hierarchical or vertical (Mountcastle, 1978; Edelman, 1987).

It is foreseeable that with non-invasive methods the brain will be increasingly better understood. At this stage, biological-integrated robots will be used for the study of memory, other neurological processes and diseases. Due to its complexity, scientists have only been able to replicate some parts that we can electronically understand, outside the body (Nili et al., 2015). Nili et al. have recently reported the first electronic multi-state memory cell, giving information of multiple processes. Once compared with actual human memory, this replica could overcome human memory capacity. In the brain, we have simultaneously old personal memories (episodic memory) and declarative memory (explicit memory of facts and events). This system is inspired in the human brain in the sense that the "electronic long-term memory cell can mimic the way the human brain processes information" (Joshi, 2015). Nili et al. (2005) suggest that the human brain and the ionic brain will be similar.

Nowadays, the *hard problem* of consciousness is shared between understanding the human brain and the emerging machine paradigms: what is the nature of experience? What is the nature of the social mind and the networked systems? An interesting example to study these phenomena is physical pain. Subjectively, pain can be programmed in a bionic brain (Joshi, 2015). It is not yet possible for a robot to feel pain in the same way that humans do, because of physiological differences. The ideas of Epicurus (341–270 BC), Descartes (1596-1650), Condillac (1715-1780) and La Metrie (1709-1751) emphasized that differentiated systems exist around us at different times of our lives (little bits of matter as *atoms*, sensitive statues, and other things). Mathematician Thomas Hobbes (1588-1679) said: "thinking is calculating". Descartes associated animals to machines, and Leibniz (1646 –1716) made a design of a reasoning machine to solve differences of "opinion" (beliefs) (Leibniz, 1685).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The last turning point of this text is the discussion of common ideas about the power of facts/fictions in consciousness, what could be called *factions*. A common *faction* is a complex system such as a machine that causes divergence and deviance, an organism that is capable of causing "damages" to the human being while doing it *consciously*.

Recent blockbuster filmography has initiated a public discussion on the following questions. Is it possible that a person possesses beliefs or feelings for an operative system (OS) or an artificial intelligent (AI) machine? Can a machine have and *cause* mental states in someone else?

Consciousness is an essential process to motivate oneself to feel that other person is capable of acting upon oneself. It was suggested that actions can "sneak up" without sufficient intent on our part (Wegner, 2002). Actions can become unpleasantly inconsistent compared to previous intentions (e.g. to create an emotional machine), therefore to urge an action that creates a new intention (e.g. to create a beautiful smile).

The AI field is working on "adapted" mental states that can be produced by other system besides humans. A small robot-cockroach learned how to behave and be accepted as a group member, in contact with other cockroaches. The aim is not to compare the robot-cockroach to the human brain, which is evolutionary and has neuroplasticity. Instead, the aim is to push the limits of self-learning computation.

In the realm of science fiction cinema, movies such as *Her* (Jonze, 2013) or *Ex machina* (Garland, 2015) suggest a different future of OS and AI machines. These movies question if it will ever be possible for a (sensitive) person to share a conscious experience with a purchased OS or an AI machine. In *Her*, an OS was the "personality" in the voice of *Samantha* (Scarlett Johansson). Voice is something ineffable and ephemeral, as thinking. In the movie the main character *Theodore* (Joaquin Phoenix) is asked by his friend: "what do you like *more* to see in *Samantha?*" What is the source of love in the relationship? Is the vocal enchantment that captivates him? It is difficult to think about the subtleties of our understanding of each other.

Norbert Schwarz et al. (2009) called for the concept of "fluency effect" in information (a metacognitive experience) which is difficult to comprehend, as it affects the information *substance*.

The idea of exchanging messages with a virtual entity (without a body) is also portrayed in OS character *Samantha*. Nowadays, with the explosion of online dating services (also portrayed in the movie), these questions curb the user's brain, since the user does not know if the person they are interacting with is in fact a human being, because they have never actually met previous to the virtual encounter.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Since approximately one third of the brain is specialized in vision (Eagleman, 2011), perception has a tendency to be ambiguous. What we see is a refugee of belief in that particular help context, however there are some tricks and assumptions. Alex Garland (director of the movie *Ex machina*, 2015) foresees these complex dilemmas, since he makes us imagine a "real" AI (Ava) that can pass the abysmal gap of achieving consciousness. The relationship based on meeting Ava distinguished itself from the scientific domain, originally dedicated to knowledge of AI, not feeling of AI. In *Ex Machina*, the willpower and knowledge of the *other* give Kaleb (Domhnall Gleeson) the feeling of self-knowledge.

This process of conscious analysis follows Friedrich Hayek thought "classification processes" or self-organization: "much of what we think we know about the outside world is indeed knowledge of ourselves" (Hayek, 1952b).

Ava (Alicia Vikander) becomes a "conscious" interlocutor because Ava is "connected", therefore conscious. Movie character of creative computation guru *Nathan* (Oscar Isaac) asks *Kaleb* at the beginning of the movie: "the challenge is to show that it [Ava] is a robot. And see if you still feel that it is aware (...)". Intentionality, this is the first dimension for a "theory of mind" (TM), when a person debates about dispositions, beliefs or desires, and Ava is there to do so. Other dimensions of intentionality are to know about *the other*, and IA is adapted to complex human relations, including verbal and non-verbal communication.

Hayek said "We cannot discard, but only develop what we do not understand" (Hayek, 1979). To feel loved, hated or betrayed, there can be delusions or illusions, such as a schizophrenic delusion that may let us believe that the brain has been exchanged with another one, such as with Ava.

Why is consciousness central? Perhaps it is to anticipate upon other people's minds, and fight against *bad* operations of survival of Ava. What about the sensitivity of feeling that defines the "explanatory gap" of consciousness? The brain has the mind and the body at various levels. However, what characterizes psychosis is the feeling of powerlessness to control daily life events. For its part, the subjective experience of a depressive person could be the appearance of life, as a gap, or a fragmentation of unconnected events.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Final Remarks

"Emotions are a collection of unconscious neural responses to *qualia*."
—Damásio

There may be a large gap between implicit or emotional knowledge and awareness.

Mind and culture are developed concomitantly and not successively (Hayek, 1979). Distinct forms of consciousness rest on other prior conditions: unconscious mental schemes are therefore abstract cognitive structures that generate experience, by recurrent models (Neisser, 1987b), or by themes about advanced knowledge on a given subject. The brain accepts *schemata*, "schematic structures". These schemes are not represented in the mind, but in the body (Johnson, 1987).

The shifting turning points of knowledge about consciousness are outlined and discussed in the light of different times and approaches. From the sensible receptivity, we think of a literary work and a philosophical one (Churchland, 2008): How to live in terms of "the experiential self"?

Qualia is a term philosophy of mind uses to refer to the mental states of senses. This refers to smell, colors or sounds. From daily experience, as soon as a person wakes up, a real sensory experience is felt (a phenomenon). This can be the smell of coffee, without the conscious process.

The turning points represent scientific moments of comprehension and fictional motivated constructs to explain consciousness. Below are summarized the main key points: (1) functional biopsychology – split-brain and human functionality – a new reality of split-brain, with *two minds* in a brain; illusion of the two minds – corpus callosum connects them; (2) integrated neuroscience – the *extended conscious* and an (unconscious) neural reaction to a certain stimulus – the emotion level; illusion of the cortex command – connection between cortical and subcortical areas; (3) emotions realm – two types of emotions – conscious (cortex) and automatic/unconscious (amygdala), and the (still unconscious) sensing of this body state (feeling); illusion of super control and body power; (4) the new unconscious – multiple people with a complex ensemble of neural activations in their brains – multiple talking minds – and feelings; illusion of mutual understanding; and (5) the step to conscious machines – two systems and complex shared feelings – the "simulator" hypothesis – a "shared activation mechanism" is needed, at a motor and intentional level, beyond the "theory of mind"; psycho-techno-thrillers illusion – multiple minds and enlarged dimensions are aware and sensible of multiple inputs and outputs.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

There is a large gap between what a person knows but does not know how is known (implicit knowledge) and what that person is aware of. In the paper "Attention alters the visual plasticity", Gutnisky et al. (2009) showed that the brain absorbs what is seen, like Ava, the character from the movie *Ex Machina*, by Alex Garland, does with consciousness.

Firstly, in order to make those cerebral psycho-techno-thrillers alive or understand them, we must grasp some mental criteria, as said by Eagleman (2011): "All vision is illusion". What the brain creates is a mental script of things. Neuroscience social knowledge reflects upon other illusions, which we thought were memories (Schacter, 1987) or "realities" of perception, more than realities as we see them. Our brain "creates" the experience that we thought of as a sensory perception. Here it is important to remember Immanuel Kant's (1724-1804) *a priori* concepts, or George Berkeley's (1685-1753) old idea: "to be is to be perceived" (*esse est percipi*). The discrepancy between what our brain registers and what we see (think and remember) is tremendous. Hence, a lot of information is always lost from the sensory system.

Secondly, the brain "waking consciousness" is always active, whether we are asleep or awake and it is difficult to know what is "normal", as we make quick decisions about mindfulness. We must be aware of the human limitations that lead us to dichotomous thought categories, such as "to be awake", when sleep is not opposed to be aware. When we think of the abstraction of a linguistic idea such as "to become awake", it is a "surface structure" of language. In that structure, transmitting the idea of being awake (e.g. by saying that a person is in possession of a *normal consciousness* or a *waking consciousness)* is not sufficient to transmit that information (tactile, unconscious or implicit). Since we enjoy the awakened experience, we have modes of storytelling about our existence and about *what happens*. That serves to the construction of the self and the world. To think on the waking process brings concrete facts and imagination to consciousness – fantasies, desires, with emotional and sometimes intense internal reactions.

Thirdly, conscious knowledge does not refer to "fatal visions" – such as ASC, as in *Macbeth* (Shakespeare, 1623) – "Art thou not, fatal vision, sensible / To feeling as to sight? / Or art thou but / A dagger of the mind, a false creation, / Proceeding from the heat-oppressed brain?" Can a sudden emotion that is terrifying, as that one, block the most *lucid dream*, and does his frontal cortex diminish the shock of emotion? His mind would wander round in hallucination, when interpretations are already erroneous behaviors. Wouldn't Shakespeare be concerned before he understands the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

interpretation of what his *other mind* would do? Would that *other mind* wonder about the hallucination of his first mind?

Finally, consciousness is seen with a social neuroscience perspective and the focus on relationship, proximity and intimacy, with not "easy" gender-neutral questions (Fitzpatrick, 2012). The network between *us* and *them*, the "family resemblance" (Wittgenstein, 1997) appears to be fundamental to realize how the categorization process happens. Representation of human characteristics increases with increasing similarities between people and no bizarre entities.

References

Amorapanth, P., LeDoux, J., & Nader, K. (2000). Different lateral amygdala outputs mediate reactions and actions elicited by a fear-arousing stimulus. *Nature Reviews Neuroscience*, *3*, 74-79.

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

—. (2002). The conscious accept hypothesis: Original and recent evidence. *Trends in Cognitive Science*, *6*, 47-52.

Baptista, A. A. (1985). *Os nós e os laços*. Lisboa: Presença.

Barbur, J., Watson, J., Frackowiak, R., & Zeki, S. (1993). Conscious visual perception without V1. *Brain, 116,* 1293-1302.

Bargh, J. (2007). *Social psychology and the unconscious: The automaticity of higher mental processes*. N.Y.: Psychology Press.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: The MIT Press.

—. (1997). Is there a language of the eyes? *Visual Cognition*, *4*(3), 311-331.

—. (2011). *Zero degree of empathy*. London: Penguin.

Becker, C. (1992). *Living and relating: An introduction to phenomenology*. London: Sage.

Bennett, M., Dennett, D., Hacker, P., & Searle, J. (2007). Neuroscience and philosophy: *b*rain*, m*ind, and language*. N.Y.: Columbia University Press.*

Berlin, H. (2011). The neural basis of the dynamic unconscious. *Neuropsychoanalysis*, *13*(1), 5-31.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

—. (1991). *The narrative construction of reality*. *Critical Inquiry*, *18,* 1-21

—. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Carey, B. (2008). Blind, yet seeing. The brain subconscious visual sense. *The New York Times*, Dec. 23, 2008.

Carter, R. (1998). *Mapping the mind*. London: Weidenfeld & Nicolson.

Celesia, G. (2010). Visual perception and awareness: A modular system. Journal of Psychophysiology, 24(2), 62–67.

Churchland, P. (2008). The impact of neuroscience on philosophy. *Neuron 60*, nov. 6, 409-411.

Cowen, P., Harrison, P., & Burns, T. (2012). Shorter Oxford Textbook of Psychiatry (6th Ed.). Oxford: Oxford University Press,

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. N.Y.: Scribner.

Crick, F. & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature, 375,* 121-123.

Crick, F. & Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex*, *8*, 97-107.

Damásio, A. (1994). *Descartes' Error: Emotion, reason and the human brain*. London: Piscator.

—. (1999). *The feeling of what happens: Body, emotion in the mankind of consciousness*. N.Y.: Harcourt.

Damásio, H., Grabowski, T., Tranel, D., Hichwa, R., & Damásio, A. (1996). A neural basis for lexical retrieval. *Nature,* 380 (6574), 499–505.

Davidson, R. & Begley, S. (2012). *The emotional life of your brain*: *How its unique patterns affect the way you think, feel, and live--and how you can change them*. London: Penguin.

Doidge, N. (2007). *The brain that changes itself: stories of personal triumph from the frontiers of brain science*. London: Penguin.

Drevets W., Price, J., Simpson, J. Jr, Todd R., Reich, T., Vannier, M. &, Raichle, M. (1997). Subgenual prefrontal córtex abnormalities in mood disorder. *Nature*, 386, 6527, 284-287.

Eagleman, D. (2011). *Incognito: The secret lives of the brain*. N.Y.: Pantheon.

Edelman, G. M. (1987). *Neural darwinism: The theory of neuronal group selection*. N.Y.: Basic Books.

Edelman, G. & Tononi, G. (2000). *Consciousness: How matter becomes imagination*. N.Y.: Basic Books.

Fletcher, P., Happé, F., Frith, U., Baker, S. C., Dolan, R., Frackowiak, R., & Frith, C. (1995). Other minds in the brain: a functional imaging study. *Cognition*, *57*, 109-128.

Fridja, N. (2007). *The Laws of emotions.* N.J.: Lawrence Erlbaum Association.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Frith, U. (1989). *Autism: Explaining the enigma*. Oxford: Blackwell.
Fitzpatrick, S. (2012). Functional brain imaging: Neuro-turn or wrong turn? In M. Littlefield & J. Johnson (Eds.), *The neuroscientific turn*: *Transdisciplinary in the age of the brain* (pp. 180-198). Ann Arbor: University of Michigan Press.
Furnham, A. (2008). *50 Psychology ideas you really need to know*. London: Quercus Editions.
Galbis-Reig, D. (2004). Sigmund Freud, MD: Forgotten contributions to neurology, neuropathology, and anesthesia. *Internal Journal of Neurology*, *3*(1).
Gardner, H. (1986). *The mind's new science: A history of the cognitive revolution.* N.Y.: Basic Books.
Garland, A. (2015). Director of *Ex Machina*, Produced by DNA Films.
Gazzaniga, M. (1967). Split brain in man. *Scientific American*, *217*(2), 24-29.
—. (1992). *Nature's mind: The biological roots of thinking, emotions, sexuality, language and intelligence*. Harmondsworth: Penguin Books.
—. (1995). Consciousness and de cerebral hemispheres. In M. Gazzaniga (Ed.), *The cognitive neuroscience* (pp. 1391-1400). Cambridge, MA: MIT Press.
—. (1998). The split brain revisited. *Scientific American*, 279, 1 (Jul. 1998). 51-55.
Graziano, M. (2013). Conscious and the social brain. Oxford: Oxford University Press.
Gutnisky, D., Hansen, B., Iliesen, B. & Dragoí, V. (2009). Attention alters visual plasticity during exposure-based learning. *Current Biology*, *19*(7), 555-560.
Hassin, R., Uleman, J., & Bargh, J. (2005). (Eds.). *The new unconscious*. Oxford: Oxford University Press.
Hayek, F. (1952b). *The sensory order*. Chicago: University of Chicago Press.
—. (1978). *New studies in philosophy, politics, economics, and the history of ideas.* Chicago: University of Chicago Press.
James, W. (1884). What is an emotion? *Mind*, *9*(39), 188-205.
Jonze, S. (2013). Director of *Her*, Annapurna Pictures
Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination and reason*. Chicago: University of Chicago Press.
Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.
Joshi, M. (2015). Can we experiment on a bionic brain if it feels human pain? On website: http://bigthink.com/ideafeed/roboethics-can-we-

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

experiment-on-a-bionic-brain-that-is-capable-of-human-pain (accessed at 27 Mars 2015)

Kaas, J. H. (2006). Evolution of the neocortex. *Current Biology*, *21*(16), 2006, R910-914.

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., & Hudspeth, A. (2000). *Principles of neural science* (4th. Ed.). N.Y.: McGraw-Hill.

Keenan, J., Nelson, A., O'Connor, M., &, Pascual-Leone, A. (2001). Self-recognition and the right hemisphere. *Nature*, *409*, 305.

Kihlstrom, J., Barnhardt, T., & Tataryn, D. (1992). The psychological unconscious: Found, lost, and regained. *American Psychologist*, *47*(6), 788-791.

Koch, C. (2009). "Minds, brains and society" (lecture in Caltech, Pasadena, CA, 21 Jan. 2009).

—. (2012). *Consciousness: Confessions of a romantic reductionist.* Massachusetts: The MIT Press.

Kolb, B. & Whishaw, I. (2004). *An introduction to brain and behaviour*. N.Y: Worth.

Kosslyn, S. & Rosenberg, R. (2004). *Psychology: The brain, the persona, the world* (2nd. Ed.). N.Y.: Pearson.

Kringelbach, M. (2005). The human orbitofrontal cortex: Linking reward to hedonic experience. *Nature Reviews: Neuroscience*, *6*(9) 2006, 691-702.

LeDoux, J. (1996). *Emotional brain: The mysterious underpinning of emotional life.* N.Y.: Simon & Schuster).

—. (2012). Rethinking emotional brain. *Neuron*, *73*, 653-676.

LeDoux, J., Wilson, D., & Gazzaniga, M. (1977). A divided mind. *Annals of Neurology*, *2*, 417-421.

Leibniz, G. (1685). *The art of discovery.* In P. Weiner (Ed.), *Selections* — Gottfried *Wilhelm Leibniz.* N.Y.: Charles Scribners.

Leopold, D. & Logothesis, N. (1996). Activity changes in early visual cortex reflect monkeys' precepts during binocular rivalry. *Nature*, 379-549-553.

Lerner, M. (1980). The belief in a Just World*: A fundamental delusion*. N.Y: Plenum.

Leslie, A. (1991). The theory of mind impairment in autism: Evidence for a modular mechanism of development. In A. Whiten (Org.), *Natural theories of mind* (pp. 63-78). Oxford: Blackwell.

Levin, D. & Simons, D. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, *4*(4), 501-506.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Lillard, A. & Skibbe, L. (2004). Theory of mind: Conscious attribution and spontaneous trait inference. In R. Hassin, J. Uleman, & J. Bargh (Eds.), *The new unconscious* (pp. 277-308). Oxford: Oxford University Press.

Llinas, R. R., Ribary, U., Joliot, M., & Wang, X.-J. (1994). Content and context in temporal thalamocortical binding. In G. Buzsaki, R. R. Llinas, & W. Singer (Eds.), *Temporal coding in the brain*. Berlin: Springer Verlag.

Loftus, E. (1974). Reconstruction of automobile destruction – Example of interaction between language and memory. *Journal of Verbal Learning and Verbal Behaviour, 13*(5), 585-589.

—. (2005). Planting misinformation in the human mind: S 30-year investigation of the malleability of memory. *Learning & Memory*, *12*, 361-366.

Logothetis, N. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869-878.

Maslow, A. (1954). *Motivation and personality*. N.Y.: Harper.

Miguéis, J. R. (1958). *Leáh e outras histórias*. Lisboa: Editorial Estúdios Cor.

Mlodinow, L. (2012). *Subliminar: How your conscious mind rules your behaviour.* N.Y.: Vintage Books.

Mountcastle, V. (1978). An organizing principle for cerebral functions: The unit module and the distributed system. In G. Edelman & V. Mountcastle (Eds.), *The mindful brain* (pp. 1-50). Cambridge MA: The MIT Press.

Naselaris, T., Prenger, R., Kay, K., Oliver, M., & Gallant, J. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron 63*, 902-915.

Nass C., Tauber J., & Reeves E. (1994). Computers are social actors. Proceedings of CHI'94: Human Factors in Computing Systems, 72-77. Boston, MA, Association for Computing Machinery.

Nass C., Moon, Y., & Green, N. (1997). Are computer gender neutral? *Journal of Applied Social Psychology*, *27*(10), 864-876.

Natsoulas, T. (2001). On the intrinsic nature of states of consciousness: Attempted inroads from the firs-person perspective. *Journal of Mind and Behaviour*, *22*, 219-248.

Negroponte, N. (1989). From Bezel to Proscenium. In Proceedings of SigGraph '89.

Neisser, U. (1987b). *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge: Cambridge University Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Nili, H., Walia, S., Kandjani, A., Ramanathan, R., Gutruf, Ph., Ahmed, T., Balendhran, S., Bansal, V., Strukov, D., Kavehei, O., Bhaskaran, M., & Sriram, S. (2015). Donor-Induced Performance Tuning of Amorphous $SrTiO_3$ Memristive Nanodevices: Multistate Resistive Switching and Mechanical Tunability. *Advanced Functional Materials*, April 14, 2015.

Nir, Y. & Tononi G. (2010). *Trends in Cognitive Sciences*, *14*(2), 88-100.

Northcutt, R. & Kaas, J. (1995). The emergence and evolution of mammalian neocortex. *Trends of Neuroscience*, *18*(9), 373-379.

Ochsner, K. & Lieberman, M. (2001). The emergence of social cognitive neuroscience. *American Psychologist 56*(9), 717-728.

Parkin, A. (1996). Explorations in cognitive neuropsychology. Oxford: Oxford University Press.

Payne, D., Elice, C., Blackwell, J., & Neuschatz, J. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Learning & Memory*, *35*, 261-285.

Pegna, A., Khateb, A., Lazeyras, F., &, Seghier, L. (2005). Discriminating emotional faces without primary visual cortices involves the right amygdala. *Nature Neuroscience, 8*(1), 24-25.

Penrose, R. (1989). *The emperor's new mind.* Oxford: Oxford University Press.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behaviour and Brain Sciences*, *4*, 515-526.

Rakic, P. (2010). The evolution of neocortex: Perspective from developmental biology. *Nature Reviews Neuroscience*, *10*, 724-735.

Rosenthal, D. (2005). *Consciousness and mind.* Oxford: Oxford University Press.

Rossano, M. (2003). Expertise and the evolution of consciousness. *Cognition*, *87*(3), 207-236.

Schwarz, N., Song, H., & Xu, J. (2009). When thinking is difficult: Metacognitive experiences as information. In M. Wänke (Ed.), *Social Psychology of Consumer Behaviour* (pp. 201-223). N.Y.: Psychology Press.

Schacter, D. L. (1987). Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *13*, 501-518

Shakespeare, W. (1623). *Macbeth*. London Penguin (Version of 1988).

Silvia, P. J. (2002). Self-awareness and emotional intensity. *Cognition & Emotion*, *16*, 195-216.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Simons, D. & Levin, D. (1989). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*(4), 644-648.

Singer, W. (1998). Consciousness and the structure of neuronal representation. *Philosophical Transactions of the Royal Society Britannica*, *353*, 1829-1840.

Solms, M. (2004). Freud returns. *Scientific American*, *5*, 83-88.

Sperry, R. W. (1968). Hemisphere disconnection and unity in conscious awareness. *American Psychologist*, *23*, 723-733.

—. (1974). Hemispheric specialization: scope and limits. In F. Schmitt & F. Worden (Eds.), *Neuroscience: Third Study Program* (pp. 5-19). Cambridge, MA: The MIT Press.

Springer, S. & Deutsch, G. (1993). *Left brain/right brain* (4th. Ed.). N.Y.: W. H. Freeman.

Striedter, G. F. (2005). *Principles of brain evolution.* Sunderland, MA: Sinauer Associates.

Tiberghien, G., Abdi, H., Desclés, J.-P., Georgieff, N., Jeannerod, N., Le Ny, J.-F., Livet, P., Pynte, J., & Sabah, G. (2002). *Dictionnaire des sciences cognitives*. Paris: Arman.

Varela, F. & Shear, J. (1999) (Eds.). *The view from within: First person methodologies*. London: Imprint Academic.

Yasnitsky, A. (2009). Ocherk istorii Khar'kovskoj psikhologicheskoj shkoly: pervaya nauchnaya sessiya Khar'kovskogo gosudarstvennogo instituta i poyavlenie "Khar'kovskoj shkoly psikhologii" (1938) [An outline of the history of the Kharkov school: first scientific session of the Kharkov state pedagogical institute and the emergence of the "Kharkov school of psychology" (1938)]. *Cultural-Historical Psychology* (2), 95–106.

von der Malsburg, C. (2002). How are neural signals related to each other and to the world? *Journal of Consciousness Studies*, *9*, 47-70.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral cortex*, *21*, 2461-2470.

Wegner, D. (2002). More than good intentions: Holding fast to faith in free will. *The New York Times*, 31 Dec. 2002.

Weiskrang, L. (1986). *Blindsight: A case study and its implications*. Oxford: Clarendon Press.

Welborn, B. & Lieberman, M. (2015). Person-specific Theory of Mind in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 27*, 1-12.

Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Belknap Press, 5.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Wittgenstein, L. (1997). *Philosophical investigations.* Oxford: Blackwell (Originally published 1953).

Zamith-Cruz, J. (1996). Trajectórias criativas: O desenvolvimento humano na perspectiva da psicologia narrativa. Unpublished doctoral dissertation, Universidade do Minho – Portugal.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

PART V:

PHILOSOPHY OF MIND: HISTORY, INFLUENCES AND CONCEPTS

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER FOURTEEN

PRIVILEGED ACCESS TO CONSCIOUS EXPERIENCE AND THE TRANSPARENCY THESIS

KLAUS GÄRTNER

I. Introduction

Epistemology in philosophy of mind is a difficult endeavor. Consider for example the more than problematic claim about our special epistemic relation with our own mental states. Those who believe that our conscious experiences are different from other domains suggest that self-knowledge about phenomenal properties is certain and therefore privileged. Usually, this so called privileged access is explained by the idea that we have *direct* access to our phenomenal life. This means, in contrast to perceptual knowledge, self-knowledge is non-inferential. It is widely believed that this kind of directness involves two different senses: an epistemic sense and a metaphysical sense. Proponents of this view often claim that privileged access is an important folk psychological intuition. As a consequence, they hold that introspection is different from perception. Unfortunately this approach has to deal with a serious objection stemming from the claim that experiences are transparent.

At the beginning of the last century, G.E. Moore in his attempt to refute idealism[1] introduced the intuition that experiences are diaphanous and changed the way we think about the ontology of experience in a radical fashion. His keen observation basically means that in perception we usually do not become aware of conscious features of an experience, we rather become aware of features of the objects those experiences are about. Moore's conclusion however was modest, he simply thought that mind-independent objects exist, concluding that the primary target of

[1] See Moore, 1903.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

introspection are the objects of experience, not the intrinsic features of consciousness.

However, it was not until Gilbert Harman that the transparency thesis showed its real potential. Harman's thesis[2] claims that there is nothing more we can know about experiences than the features of the intentional objects, reducing conscious features to purely representational ones. If true, this leaves no room for privileged access. Since introspection and perception are exhausted by the experience's representational features, there can be no essential difference in the gained knowledge. Despite the fact that many opponents of representational theories of mind deny Harman's strong conclusion, most of them admit that transparency affects the nature of conscious experience, maintaining however the existence of privileged access.

Because of the tremendous implications of the transparency thesis on conscious experience in general, I want to show how it influences the debate about privileged access. Firstly, I will give a short overview of both intuitions. Secondly, I will state – what I think are – the two main ideas of how transparency enters the privileged access discussion. Thirdly, I will briefly review – what I consider – the most important views in this context. And finally, I will give a short road map of what has to be done to satisfy both intuitions adequately.

II. Privileged Access

As introduced above, privileged access captures the epistemological specialness of self-knowledge about our own mind. According to Gertler this means the following: "Self-knowledge may be epistemically special in that (a) it is especially secure or *certain*; (b) one uses a unique method to determine one's own mental states"[3]. Of course both epistemologically special characteristics are not exclusive.

Let me start with (a). In the case of self-knowledge the epistemically strongest ideas are infallibility and omniscience. Gertler explains these claims in the following way:

> One is *infallible* about one's own mental states if, and only if, one cannot have a false belief to the effect that one is in a certain mental state. (In other words, one's belief that one is in a particular mental state entails that

[2] See Harman, 1990 and 1996.
[3] Gertler, 2015, § 1.1.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> one is in that mental state.) One is *omniscience* about one's own states if, and only if, being in a mental state suffices for knowing that one is in that state. (In other words, one's being in a particular mental state entails that one knows that one is in that state).[4]

Both claims are particularly strong. As a consequence, nowadays hardly anyone thinks this to be true.

Restricting those ideas means basically limiting their scope. Not all beliefs about our own mental states are infallible or omniscient, only the ones formed by the special method of introspection. We could put the weaker thesis as follows: "when one carefully, attentively employs the mode of knowing unique to self-knowledge, one will not form a false belief about one's own states"[5]. This might be problematic for all kinds of mental states, but at least for our current phenomenal states[6] or properties this seems to be true. Of course those claims can be weakened even further, but for our purposes this short characterization is sufficient.

Gertler also states another important idea. She says that "[...] infallibility and omniscience correlate the belief that p with p itself. But they are neutral between epistemic internalism and externalism"[7]. While versions of epistemic externalism speak about infallibility and omniscience as the highest degrees of epistemic security, the highest degree of epistemic security in epistemic internalist models is certainty. "The claim that one can be *certain* that one is in a particular mental state applies to a single self-attribution, whereas the reliability-based theses of infallibility and omniscience concern a person's general accuracy." [8] Epistemic certainty is often tied to the idea of introspection as a special unique method of obtaining knowledge about our own mental states[9]. Still, there are stronger and weaker versions of both theories.

Now, let me turn to (b). When we talk about the unique epistemic method to grasp one's own mental states we talk about introspection. In this particular case, we talk about introspection from an epistemic point of view. So, what makes introspection so special? According to Gertler,

---

[4] Gertler, 2011, pp. 61-62.
[5] Gertler, 2015, § 1.1.1.
[6] 'phenomenal states' is how Gertler puts it. I want to note that this is far from clear. If something phenomenal is realized as a state is controversial.
[7] Gertler, 2015, § 1.1.1.
[8] Gertler, 2011, p. 65.
[9] Since I want to explore what we can know about the phenomenal, or better what is the privileged access to the phenomenal, I will assume certainty.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

"[o]ne standard answer to this question is that we have epistemic access to our states that is *direct*, whereas our access to facts or objects external to us is indirect"[10]. This directness can come in two forms:

> In the first, epistemic sense, the claim is that we can grasp our own mental states without inference; we need not rely on reasoning from observation. The second sense of directness is metaphysical: there is no state or object that mediates between my self-attributing belief (that I am now thinking that it will rain, feeling thirsty, etc.) and its object (my thought that it will rain, my feeling of thirst).[11]

The standard approach to explain privileged access to our phenomenal life is the unmediated observation model.[12]

The unmediated observation model – often attributed to Descartes – often claims that we are acquainted with our phenomenal properties. This means that this approach holds that there is a *direct* access to a given phenomenal property; i.e. that there is no mediating state and that the knowledge obtained is non-inferential. According to Gertler, these so called self-presenting properties imply certain psychological and epistemic characteristics:

> Specifically, (i) no one who has a self-presenting property *directly* self-attributes its negation […]; (ii) anyone who has a self-presenting property and considers whether she does, will self-attribute that property; and (iii) a direct attribution of a self-presenting property is *certain*, in the relative sense.[13]

---

[10] Gertler, 2011, p. 65.
[11] Gertler, 2015, § 1.1.2.
[12] This does not mean that there are no other accounts. Proponents of the Inner-sense model, who hold that introspection is analogous to perception, can also explain privileged access to some extent. They will claim that self-knowledge about phenomenal properties is more secure in degree. This may be the case because we use certain abilities to obtain knowledge about this restricted class of mental states, namely our own, but they "[...] will deny that the difference between self-knowledge and other types of knowledge have deep philosophical significance" (Gertler, 2011, p. 66). For reasons of space, I will not discuss this possibility. I, however, think that, since proponents of this view already assume that privileged access has no 'deep philosophical implications', this account does not take the privileged access intuition serious.
[13] Gertler, 2015, § 2.1.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

In my opinion, defending a robust account of privileged access means to defend some version of this view.

III. Transparency of Experience

As stated in the introduction, transparency was famously presented by G. E. Moore in his article 'The Refutation of Idealism' (1903). In the original context, Moore suggested that perceptual experience is diaphanous to challenge the idealist's claim that objects of experience are mind-dependent. According to the general idea we do not become aware of the conscious features of our experiences, we rather become aware of features of the objects – which do not depend on consciousness itself – those experiences are about. To put it in Moore's words:

> [...] the moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation blue, all we can see is the blue: the other element is as if it were diaphanous.[14]

According to this initial claim, it is not necessarily the case that features of consciousness do not exist. It rather means that there is a unique relation to those mind-independent objects presented by experience. Moore therefore separates two elements of experience, namely what we today refer to as the experience's content and its phenomenal properties. A contemporary theory in Moore's spirit therefore claims that whether or not phenomenal properties really entail phenomenal blueness is difficult to say; introspection, however, refers to the representational content 'blue'.

Certainly, Moore's original thesis had its own impact in the history of philosophy of mind[15], but it was not until Gilbert Harman[16] that it reached its full potential. Harman picked up the argument and implemented it in the contemporary debate. His radical interpretation of the structure of experience and the relation to the objective world are the corner stone to Representationalism and Intentionalism.

Now, Harman's analysis constitutes a stronger interpretation of transparency, claiming a more radical consequence from the diaphanousness

[14] Moore, 1903, pp. 21-22.
[15] See e.g. Broad, 2009; Ryle, 1949; and Grice, 1961.
[16] See Harman, 1990 and 1996.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

of experience. Consider Harman's example of a red ripe tomato. Describing the situation he concludes the following:

> When you think about visual representation, it is very important to distinguish (A) qualities that the experience represents the environment as having from (B) qualities of experience by virtue of which it serves as a representation of the environment. When you see a ripe tomato your visual experience represents something as red. The redness is represented as a feature of the tomato, not a feature of your experience.[17]

To make his point, Harman does not only stress that the red feature is represented as being in the world, but it is also the case that one cannot know whether or not one's experience entails phenomenal redness. Since one is not in a position to consciously access those phenomenal properties, one cannot know anything about them at all. Introspection, therefore, fails to tell us something about those properties. Due to this fact, the only way one can obtain the concept of red is by abstracting from red objects in the world. The reason is that, according to Harman, there is a vital distinction between properties of the object of experience and properties of the experience of an object[18]. He explicitly denies conscious access to the latter, making phenomenal qualities nothing other than representational qualities. Especially Representationalists and Intentionalists still support this version of transparency[19]. (Of course, it also attracts many critics.[20]) In the general context of this debate, it is important to understand that any thesis that subscribes to the strong transparency claim must at least entail the following two assumptions:

(1) By introspection, one becomes aware of mind-independent objects of experience.
(2) Introspection constitutes no awareness of intrinsic features of experience at all.[21]

---

[17] Harman, 1996, p. 8.
[18] See Harman, 1990.
[19] See e.g. Tye, 1995, 2000, 2009; Martin, 2002; and Byrne, 2001.
[20] See e.g. Block, 1990 and 1995; Nida-Rümelin, 2007a and 2008; and Stoljar, 2004.
[21] These claims are taken from Crane, 2014.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

IV. Privileged Access and Transparency

1. Two Relations

But what happens when both intuitions about our ongoing conscious experiences collide? Transparency is an epistemic thesis about experiences. As stated in the previous section, it basically claims that introspection refers to what experiences are about. As a consequence, it influences the epistemic condition of the privileged access which assumes that the justification of privileged introspective self-knowledge depends only on the subject's conscious state. This means that, if I form a judgment about an ongoing red experience for example, for that judgment to count as this kind of knowledge, it will depend only on the current red experience itself for justification.

Now, there are two possibilities in which transparency can enter the picture. The first manner leans on Harman's stronger claim. Since we are talking about knowledge of phenomenal properties, this view suggests that they form part of the experience's representational content. In this case, conscious experience justifies a judgment via its representational content. This is to say that the introspective phenomenal judgment about an ongoing red experience depends on the experience's representational content red. Since the representational content, however, depends entirely on the experience's object, the phenomenal judgment does as well. Even though this is a viable theory, anyone who defends the privileged access intuition should resist it. The problem is the following: phenomenal judgments, according to this approach, may depend solely on current conscious experiences; the phenomenal properties of those experiences, however, depend on the experience's objects. As a consequence, justification of those judgments also depends on the experience's object. According to many proponents[22] of this strong version of transparency the qualities or properties of those objects are representational in character and therefore determined externally. But this seems implausible. If this version of transparency is true and the privileged access intuition is true, then every perceptual knowledge is also privileged. Proponents of this version of transparency, therefore, deny the privileged access intuition.

There is, however, an alternative. One can maintain both elements by denying that the previous interpretation of transparency is too strong. In

[22] See e.g. Byrne, 2001; Harman, 1990 and 1996; Martin, 2002; and Tye, 1995, 2000 and 2009a. For detailed discussion see especially Jackson, 2006.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

this case, phenomenal properties do not have to form part of the representational content of experiences.[23] Without discussing possible ontologies[24] in detail, privileged access will only depend on phenomenal properties or phenomenal reality. Since the phenomenal properties potentially have an independent ontological status, judgments about those properties will only depend on them. This seems to me an appropriate and elegant solution, since all that is needed to explain epistemic specialness are conscious experiences.

Such a view however comes with a trade-off, namely its *prima facie* incompatibility with physicalism. In a similar context, Levine introduces what he calls the Materialist Constraint. This constraint states that "[…] no appeal [can] be made in the explanation to any mental property or relation that is basic."[25] This means two things for the proponent of the privileged access intuition. Either she bites the bullet and defends anti-physicalism or she tries to explain the special access to our conscious experiences differently.

2. Views

The first combination implies – due to the strong transparency claim – that our access to our own conscious experiences is exhausted by the experience's representational content and, therefore, qualities like, e.g., red are in the world. However, the question that arises is the following: what makes this way of thinking about experiences so attractive? One straightforward answer the transparency theorist[26] can give is that treating phenomenal qualities as qualities represented by experiences is to say that experiences have those properties. This implies that there is no mystery about the phenomenal qualities, since they are not intrinsic properties of an experience. They rather form part of the content. It has often been thought that there is something wrong with this view. Critics[27] usually claim two things: one is disapproval of the implied view of introspection; the other states that a careful analysis of 'awareness' shows otherwise.

The second combination allows for more than one interpretation. Since phenomenal properties do not have to consist in representational

[23] See e.g. Block, 1990, 1996 and 2001; Chalmers, 1996; Nida-Rümelin, 2007b and 2008; Shoemaker, 1994a, b; and Stoljar, 2004.
[24] For extensive discussion see Chalmers, 2003.
[25] Levine, 2007, p. 150.
[26] See Schwitzgebel, 2014; and Tye, 2000 and 2015 for discussion.
[27] See Broad, 2009; Nida-Rümelin, 2007a; and Stoljar, 2004 for discussion.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

properties, they have *prima facie* an independent ontological status. Especially anti-physicalists[28] are in a position to construct a strong case. Privileged access, or so they claim, is due to the fact that we are acquainted with our phenomenal properties. Being acquainted with a phenomenal property means the following:

> [Acquaintance Approach] Some introspective knowledge consists in judgments that
>
> 1) are directly tied to their truthmakers;
> 2) depend, for their justification, only on the subject's conscious states at the time of the judgment; and
> 3) are more strongly justified than any empirical judgments that do not meet conditions (1) and (2).[29]

For proponents, this relation is what is special and secures the privileged access.[30] The reason why this relation is privileged is because phenomenal properties do not form part of the content, they are of the experience. As stated above, the main problem with this view is that it treats phenomenal properties as basic and therefore mysterious.

There is, however, a way out by trying to naturalize the acquaintance relation. A promising attempt is the phenomenal concept strategy[31]. Both the physicalist and the anti-physicalist can agree that concepts of our phenomenal properties are special in the sense of the privileged access intuition. Balog, however, claims that the latter often thinks that this necessarily involves the independent ontological status of phenomenal properties. This status constitutes phenomenal concepts that are directly related to those properties via acquaintance. It is, therefore, the ontological independence of the phenomenal that accounts for the specialness of phenomenal concepts.[32] The proponent of the former view does not depend on this assumption. According to Balog, in this view there is no such thing as ontologically independent phenomenal properties. There is only '*dualism of concepts*'[33]. This last idea is, according to the physicalist, also the reason why dualism *seems* to be true. Of course, this view is

[28] See footnote 24.
[29] Gertler, 2012, p. 99.
[30] For a defense see Gertler, 1999; and Nida-Rümelin, 2015.
[31] See Stoljar, 2005 for the name.
[32] Chalmers, 2003 explains in length how *direct* phenomenal concepts depend on acquaintance. For discussion see Balog, 2009.
[33] Balog, 2009, p. 303.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

compatible with this weaker account of transparency, since it is weaker than the anti-physicalist one. It is, however, not compatible with the stronger interpretation, because it still predicts special concepts about the phenomenal. This is something the latter thesis denies.

A popular way of spelling this strategy out is the constitutional account of phenomenal concepts[34]. The specialness of this account is that it straight forwardly explains the epistemic relation to our phenomenal properties. According to Balog, "[o]n the constitutional account, tokens of a phenomenal concept that refers to a particular type of visual experience [...] are constituted in part by tokens of that type of experience"[35]. This means that tokens of a certain type of experience act as 'modes of presentation of the phenomenal properties'[36] which are instantiated by them. Balog compares the constitutional account of phenomenal concepts to linguistic quotation.

> The idea of an item partly constituting a representation that refers to that item is reminiscent of how linguistic quotation works. The referent of '—' is exemplified by whatever fills in the blank. In a quotation expression, a token of the referent is literally a constituent of the expression that refers to a type which it exemplifies, and that expression has its reference (at least partly) in virtue of the properties of its constituent. [37]

While there is an account that resembles only slightly the linguistic counterpart, prefixing the experience itself by the operator 'the experience...'[38] to produce phenomenal concepts, Balog thinks that to explain those concepts one should take the quotational analogy more seriously and focus on the conceptual role of phenomenal concepts.[39] Both versions, however, fall under the name quotational account of phenomenal concepts.[40] The latter explanation states the following:

> […] on this view, every token of a phenomenal concept applied to current experience is (partly) constituted by *that token experience*, and this fact is

---

[34] Proponents of this strategy include e.g. Balog, 2006, 2012a, b; Block, 2007; Hill and McLaughlin, 1999; and Papineau, 2002 and 2007.
[35] Balog, 2009, p. 307.
[36] Balog, 2012a, p. 7.
[37] Balog, 2009, p. 308.
[38] *Ibid.*, p. 308.
[39] See Balog, 2009 and 2012b.
[40] For a proponent of the former account see Papineau, 2002 and 2007; for one of the latter account Balog, 2012a, b.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> crucial in determining the reference of the concept. Not only is it the case that a token experience that constitutes a token phenomenal concept instantiates the phenomenal property the concept refers to, but it is *because* the concept is so constituted that it so refers.[41]

In Balog's opinion, this physicalist account of phenomenal concepts can explain the acquaintance relation in the appropriate manner. The reason is that the phenomenal concepts applied contain actual instantiations of the referent physically. Since, however, tokens of the phenomenal concepts present that referent – the experience tokens – as phenomenal properties, the reference to those properties is direct, grounding the acquaintance relation[42] and, therefore, privileged access.

The main problem of this account is that the physical structure cannot explain the cognitive structure. According to Levine[43], assuming a representational system, what is important for acquaintance or cognitive presence is the relation between cognitive property tokens and not how those tokens relate to their objects. The latter relation only determines what is represented, leaving it unclear how this representation relation can account for cognitive significance. This means that difference in the former mechanism does not explain differences in what is relevant cognitively. In short, Levine's argument undermines the constitutional account's claim that substantial cognitive presence, which explains substantial acquaintance, can be explained by physical presence, denying that the physical presence is able to account for what is cognitively relevant.

V. A Road Map

In this final section, I will point out what road we could take to justify the privileged access intuition in the light of transparency. Obviously, it should be clear after reading the last section that, in my opinion, the stronger version of the latter thesis clashes with the former intuition. Unless we are willing to accept strange outcomes – i.e. perceptual knowledge is metaphysically and epistemically direct – this pair is not suited to combine both ideas. As a consequence, I will focus on the second option, namely the combination of privileged access and a weaker interpretation of transparency.

---

[41] Balog, 2012a, p. 7.
[42] See Balog, 2012a.
[43] See Levine, 2007.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Now, the basic idea of privileged access laid out in the first section is that it is a) especially epistemically secure or certain, and b) obtained by a unique method. The standard justification, or so I argued, is via the unmediated observation model. This model describes the special method, i.e. introspection, as metaphysically and epistemically direct. It is widely believed that this is not the case for any mental state, only for currently ongoing conscious experiences and their phenomenal properties. Applying the weaker transparency thesis, in this context, means that introspection gives us primarily knowledge about the experience's representational content. But introspection analyzed in the right way – e.g. as awareness instead of inspection – may tell us something about our phenomenal properties as well. Without discussing whether or not this means that phenomenal properties really have to entail e.g. phenomenal redness[44], I will briefly describe three roads the privileged access proponent could take.

The first possible road to take is trying to defend that the acquaintance relation between privileged knowledge about the phenomenal is justified by the independent ontological status of phenomenal properties. To avoid the pitfall of basic or mystery properties, one could argue for a metaphysical description of the world that secures their status.[45] It may not be the most obvious path to take, but it is, in my opinion, a serious option.

Choosing the second road means to defend that the acquaintance relation can be justified in a physicalist framework. As shown in the last section the phenomenal concept strategy may be the way to go. To overcome Levine's challenge, one would have to claim that the relation between cognitive property tokens and their objects also determines the significance relation between those cognitive property tokens. As far as I can see, one solution may lie in exploring teleosemantics[46] or teleosemiotics[47].

---

[44] As an alternative ontology one could argue that what makes a property phenomenal is not the quality it possesses or the 'what it is like'. This line of thought usually claims that what is special to the phenomenal are the subjective properties or the 'for me'. Some believe that in this context the qualities are exhausted by the content of the experience, while the subjective is not. For detailed discussion see e.g. Gallagher & Zahavi, 2015; Goldman, 1970; and Kriegel, 2003a, b and 2004.

[45] One example is Chalmers, 2012.

[46] See e.g. Neander, 2012 for discussion.

[47] See e.g. Hutto & Myin, 2013.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

A final road to take is to abandon the acquaintance approach to justify privileged access and find an alternative.[48] In my opinion, the two best candidates are self-presentation and revelation. The former epistemic principle constitutes an alternative to the unmediated observation model. When talking about the possibility of knowledge, or what we can know, Chisholm defines self-presentation in the following way:

> If (i) the property of being-F is such that every property it conceptually entails includes the property of thinking, if (ii) a person S has the property of being-F and if (iii) S believes himself to be F, then it is certain for S that he is F.[49]

According to Gertler[50], this principle basically claims that psychological properties, which are self-presenting, refer to special epistemic and psychological features. One problem with this proposal, however, is that it weakens the certainty claim. According to Chisholm certainty is closely tied to what is reasonable for the subject to accept.[51] This may lead to especially justified judgments, it lacks however certainty in the strong sense. For the privileged access proponent this is one way to go. It is, however, important that she revises the above claim so that it only refers to phenomenal properties – also taking into account transparency's influence on the matter – and work out the details about what constitutes certainty.

The latter epistemic principle, namely revelation, asks for "an uncommonly demanding and literal sense of 'knowing what'"[52]. This demanding sense follows from the claim that by having an experience we are supposed to be in a position to know or simply know the *essence* or *nature* of that experience. The general standard notion may be put the following way:

> By having an experience E with phenomenal property Q, I am in a position to know or know that Q is F (for F is the essence of Q).[53]

---

[48] For a detailed list of how epistemic specialness can be achieved see Alston, 1971.
[49] Chisholm, 1990, p. 209.
[50] See Gertler, 2015.
[51] See Chisholm, 1976.
[52] Lewis, 1995, p. 141.
[53] Similar notions may be found in e.g. Damnjanovic, 2012; Lewis, 1995; Lihoreau, 2014; and Stoljar, 2009.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

If true, clearly revelation gives us an amazing insight of what experiences essentially consist in.

Now, there is one small problem. According to Lewis[54], from ascribing to this demanding sense of knowing what to the idea that revelation is incompatible with physicalism it is only a small step. If the proponent of revelation, however, can overcome this issue, we should suspect that this profound claim about experiences may be the way to go to explain privileged access.

VI. Conclusion

In this paper, I presented two intuitions, namely transparency of experience and privileged access. I outlined both views and showed how the former influences the discussion of the latter. In the end, I concluded that transparency and privileged access are only compatible in a certain setting, and, in this context, described ways of how to maintain the latter intuition.

Acknowledgements

Klaus Gärtner's work is endorsed by the CFCUL post-doctoral research fellowship (UID/FIL/00678/2013).

References

Alston, W. (1971): "Varieties of Privileged Access", in: *American Philosophical Quarterly, 8*: 223-241.

Balog, K. (2006): "Ontological Novelty, Emergence, and the Mind-Body Problem", in: G. Abel (ed.), *Kreativität,* 26-44, Hamburg: Felix Meiner Verlag.

—. (2009): "Phenomenal Concepts", in: B. McLaughlin, A. Beckermann & S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind,* 292-312, Oxford: Oxford University Press.

—. (2012a): "In Defense of the Phenomenal Concept Strategy", in: *Philosophy and Phenomenological Research, 84*: 1-23.

—. (2012b): "Acquaintance and the Mind-Body Problem", in: C. Hill & S. Gozzano (eds.), *New Perspectives on Type Identity: The Mental and the Physical*, 16-42, Cambridge: Cambridge University Press.

[54] See Lewis, 1995.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Block, N. (1990): “Inverted Earth”, in: *Philosophical Perspectives, 4*: 53-79.
—. (1995): “On a confusion about a function of consciousness”, in: *Behavioral and Brain Sciences*, *18*: 227–247.
—. (1996): “Mental paint and mental latex”, in: *Philosophical Issues*, 7: 19–49.
—. (2001): “How Not to Find the Neural Correlate of Consciousness”, in: J. Branquinho (ed.), *The Foundations of Cognitive Science,* 1-10, Oxford: Oxford University Press.
—. (2003): “Mental Paint”, in: M. Hahn & B. Ranberg (eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, 165-200, Cambridge: MIT Press.
—. (2007): “Max Black's Objection to Mind-Body Identity”, in: T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, 249-306, Oxford: Oxford University Press.
Broad, C. D. (2009): *Mind and its Place in Nature*, London, New York: Routledge.
Byrne, A. (2001): “Intentionalism Defended,” Philosophical Review*, 110*: 199–240.
Chalmers, D. (1996): *The Conscious Mind. In Search of a Fundamental Theory,* Oxford: Oxford University Press.
—. (2003): “The Content and Epistemology of Phenomenal Belief,” in: Q. Smith & A. Jokic (eds.), *Consciousness: New Philosophical Essays*, 220-272, Oxford: Oxford University Press.
—. (2012): *Constructing the World*, Oxford: Oxford University Press.
Chisholm, R. (1976): *Person and Object,* La Salle: Open Court.
—. (1990): “The Status of Epistemic Principles”, in: *Noûs, 24*: 209-215.
Crane, T. (2014): "The Problem of Perception", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Phil*osophy *(Fall 2014 Edition)*, URL = <http://plato.stanford.edu/archives/fall2014/entries/perception-problem/>.
Damnjanovic, N. (2012): “Revelation and Physicalism”, in: *Dialectica, 66*: 69-91.
Gallagher, S. & Zahavi, D. (2015): "Phenomenological Approaches to Self-Consciousness", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Phil*osophy *(Spring 2015 Edition)*, URL = <http://plato.stanford.edu/archives/spr2015/entries/self-consciousness-phenomenological/>.
Gertler, B. (1999): “A Defense of the Knowledge Argument”, in: *Philosophical Studies, 96*: 59-87.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. (2011): *Self-knowledge,* London, New York: Routledge.
—. (2012): “Renewed Acquaintance,” in: D. Smithies & D. Stoljar (eds.), *Introspection and Consciousness*, 93-128, Oxford: Oxford University Press.
—. (2015): "Self-Knowledge", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*, URL = <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/>.
Goldman, A. (1970): *A Theory of Human Action*, New York: Prentice-Hall.
Grice, H. P. (1961): “The Causal Theory of Perception”, in: *Proceedings of the Aristotelian Society* (*Supplementary Volume), 35*: 121–153.
Harman, G. (1990): “The Intrinsic Quality of Experience”, in: *Philosoph*ical Perspectives, *4,* Action *Theory and Philosophy of Mind*: 31–52.
—. (1996): “Explaining Objective Color in Terms of Subjective Reactions”, in: *Philosophical Issues, 7:* 1-17.
Hill, C. & McLaughlin, B. P. (1999): “There Are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy”, in: *Philosophy and Phenomenological Rese*arch, *59*: 445-454.
Hutto, D. D. and Myin, E. (2013): *Radicalizing Enactivism,* Cambridge: MIT Press.
Jackson, F. (2006): “Representation, Truth, Realism”, in: *The Monist, 89*: 50-62.
Kriegel, U. (2003a): “Consciousness as Intransitive Self-consciousness: Two Views and an Argument”, in: *Canadian Journal of Philosophy, 33*: 103-132.
—. (2003b): “Consciousness as sensory quality and as implicit self-awareness”, in: *Phenomenology and the Cognitive Sciences, 2*: 1-26.
—. (2004): “Consciousness and self-consciousness”, in: *The Monist, 87*: 185-209.
Levine, J. (2007): "Phenomenal Concepts and the Materialist Constraint" in: T. Alter and S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, 145-166, Oxford University Press.
Lewis, D. (1995): “Should a Materialist Believe in Qualia?”, in: *Australasien Journal of Philosophy, 73*: 140-144.
Lihoreau, F. (2014): “Revelation and the Essentiality of Essence”, in: *Symposion, 1*: 69-75.
Martin, M. G. F. (2002): “The transparency of experience”, in: *Mind and Language, 17*: 376–425.
Moore, G. E. (1903): “The refutation of idealism”, in: *Mind, 12*: 433–453.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Neander, K. (2012): "Teleological Theories of Mental Content", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, URL = <http://plato.stanford.edu/archives/spr2012/entries/content-teleological/>.

Nida-Rümelin, M. (2007a): "Grasping phenomenal properties," in T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, 307-349, Oxford: Oxford University Press.

—. (2007b): "Transparency of Experience and the Perceptual Model of Phenomenal Awareness", in: *Philosophical Perspec*tives, *21*: 429–455.

—. (2008): "Phenomenal Character and the Transparency of Experience", in: E. Wright (ed.), *The Case for Qualia*, 309-324, Cambridge: MIT Press.

—. (2015): "Qualia: The Knowledge Argument", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*, URL = <http://plato.stanford.edu/archives/sum2015/entries/qualia-knowledge/>.

Papineau, D. (2002): *Thinking About Consciousness*, Oxford: Oxford University Press.

—. (2007): "Phenomenal and Perceptual Concepts", in: T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, 111-144, Oxford: Oxford University Press.

Ryle, G. (1949): *The Concept of Mind*, London: Hutchinson.

Shoemaker, S. (1994a): "Self-knowledge and 'Inner-sense'", in: *Philosophy and Phenomenological Research, 54*: 249-314.

—. (1994b): *Identity, Cause, and Mind*, Cambridge: Cambridge University Press.

Schwitzgebel, E. (2014): "Introspection", in: E.N. Zalta (ed.), *The Stanford En*cyclopedia *of Philosophy (Summer 2014 Edition)*, URL = <http://plato.stanford.edu/archives/sum2014/entries/introspection/>.

Stoljar, D. (2004): "The Argument from Diaphanousness", in: M. Escurdia, R. Stanton & C. Viger (eds.), *Language, Mind and World: Special Issue of the Canadian Journal of Philosophy*, 341-390, Calgary: University of Calgary Press.

—. (2005): "Physicalism and Phenomenal Concepts", in: *Mind and Language, 20*: 469-494.

—. (2009): "The Argument from Revelation", in: D. Braddon-Mitchell and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*, 113-138, Cambridge: MIT Press.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Tye, M. (1995): *Ten Problems of Consciousness*, Cambridge: MIT Press.
—. (2000): *Consciousness, Color, and Content*, Cambridge: MIT Press.
—. (2009a): "Representationalist Theories of Consciousness", in: B. McLaughlin, A. Beckermann & S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind,* 253-267, Oxford: Oxford University Press.
—. (2009b): *Consciousness Revisited: Materialism Without Phenomenal Concepts*, Cambridge: MIT Press.
—. (2015): "Qualia", in: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philoso*phy *(Fall 2015 Edition)*, URL = <http://plato.stanford.edu/archives/fall2015/entries/qualia/>.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER FIFTEEN

APPROACHING DESCARTES' DUALISM: REDUCTIONISM OF HIS THEORY OF KNOWLEDGE

ALEKSANDAR RISTESKI

In this article I will consider Descartes' dualism as a consequence of reductionism of his theory of knowledge. My intent is to further crystallize the place of dualism in Descartes' thought. In order to do that, it is necessary to elucidate the origin and the nature of Descartes' dualism; after analyzing certain metaphysical purports that actually advocate a certain form of metaphysical monism, I'll try to depict Descartes' dualism not as metaphysical or ontological in nature, but as gnoseological. After that, I'll turn attention to the origin of dualism in Descartes' thought, by analyzing the character of reductionism as such, and then by presenting the reductive character of Descartes' theory of knowledge, which is the main cause of dualistic outcomes of that theory.

1. Introduction

In considering Descartes' mind-body dualism, we have to bear in mind that this dualism is not a sort of metaphysical dualism, but rather a *gnoseological* one. Descartes never claimed that the whole of reality consists of two principles, namely *soul* and *body*, and that keeping those two principles or *substances* as distinct can help us create a coherent and systematic rational reconstruction of the whole of reality. Quite the contrary; dualism, which appears to be a stumbling stone of Descartes' philosophy, can really make the aforementioned task even more difficult. Dualism appears as a problem, not as a starting point of Cartesian philosophy.

From a metaphysical or ontological point of view, Descartes shares similar ideas with *philosophical monism*, claiming that there is only one

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

absolute substance, upon which every finite substance depends. However, Descartes also sets up a quest to examine the properties and limits of human knowledge, relying on the faculties of the knowing subject alone. In doing so, Descartes in a way refrains from the connectivity of epistemological and ontological categories, namely, that various types of knowledge correspond to various ontological spheres. Although when classifying categories of ideas Descartes indeed asks for an "ontological nature" of the causes of the various types of ideas (Lee 112) (Descartes, Meditations on First Philosophy 27); however, he still gives primacy to gnoseological inquiry as he examines ideas *as* ideas, without primary jumping to conclusion about their ontological cause. That way, from examining the various categories of ideas, Descartes indeed projects certain ontological structure of reality; however, this is done by starting from the knowing subject alone, and without complete rational reconstruction of reality.

He *presupposes* and deliberately *pretends* that there are no such connections, and even entertains the possibility of the *deceiving spirit*, or *evil genius*, who could have made everything that appears to us to actually be an illusion (Descartes, Meditations on First Philosophy 16, 166). That way Descartes rejects any possible metaphysical foundation of knowledge, and attempts to approach the task from the opposite direction – by determining the nature of the human mind and by determining that which is absolutely and indubitably known.

However, this task was not pursued without difficulties; by uncovering the real nature of the human mind, and what can be known *clearly* and *distinctly*, Descartes has also uncovered the limits of *lux rationalis* (or of his theory of knowledge, at least). He argues that mind and body can only be known clearly and distinctly as separate substances. However, our experience, and common sense as well, tells us that those two substances are in some interaction, and hence there has to be some meeting point, or the point of their union (Cottingham, The Mind-Body Relation 181-183). Since mind and body are the concepts that exclude one another, that unity cannot be known *clara et distincta*. Descartes never argued that, since *res extensa* and *res cogitans* are known as distinct substances, there is no interaction between them, nor did he argue that those two substances are metaphysical principles of all being; principles upon which everything can be deduced and explained. Rather, he saw the dualism of his theory of knowledge as problematic, attempting throughout his life to solve this problem.

In this article, my intention is to show that Descartes' dualism is the consequence of the *reductive character* of his theory of knowledge.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Descartes uses one set of parameters attempting to explain the totality of human experience; however, it appears that there are some sets of phenomena that are not compatible with presupposed explanatory parameters. Instead of giving an all-encompassing account, as he had intention to do, Descartes' reductionist theory of knowledge has *multiplied* the world of human experience. In approaching Descartes' dualism as a flaw of his theory of knowledge, I will first turn my attention to certain metaphysical purports of Descartes' that show clear similarity with *monistic* metaphysics. This will be an attempt to show that Descartes' dualism is not metaphysical or ontological. After that, I will give a short account of the nature of reductionism, and the way reductionism turns into dualism, giving Descartes' theory of knowledge as an example. This will further demonstrate why Descartes' dualism cannot be understood as metaphysical, since the truth-criteria of *clara et distincta* at Descartes' is conceptually insufficient to serve as the basis of a complete, rounded and systematic metaphysics, as well as the basis of *mathesis universalis*. I won't however advocate any *rejection* of reductionism, but rather turn attention to its flaws.

2. Descartes' Metaphysical Monism and Gnoseological Dualism; the Relation between His Metaphysics and Theory of Knowledge

Philosophical dualism is one of the hallmarks of Descartes' philosophy. However, dualism appears there not as a solution to certain problems, but as another problem that requires an appropriate answer and philosophical reflection. Descartes' philosophy is mainly depicted as dualistic, that is, we could say that it is most influential for the problems it poses considering the mind-body dualism. In his *Meditations on First Philosophy*, among his other works, Descartes formulated the well-known dualistic stance that appears to strike not only early modern philosophy, but remains a hardly answerable problem even today (Descartes, Meditations on First Philosophy 20) (Heil 16). However, in considering this problem, some scholars seem to oversee that Descartes, in his *Principles of Philosophy*, provides something like a *monistic* view. Such a view seems to be supported by his claims that there is *only one substance* in the full meaning of the word, namely *God*; in the section where Descartes argues what the substance is and how to understand the term in relation to God and finite beings, he says:

> [T]here can be conceived but one substance which is absolutely independent, and that is God. We perceive that all other things can exist

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> only by help of the concourse of God. And, accordingly, the term substance does not apply to God and the creatures UNIVOCALLY, to adopt a term familiar in the schools; that is, no signification of this word can be distinctly understood which is common to God and them. (Descartes, The Principles of Philosophy 20)

With such a claim, Descartes' philosophy seems to be more similar to that of Spinoza, who also claims that God alone is the only absolute being or the substance. This claim also resembles certain elements similar to that of Thomistic and Aristotelian philosophy. However, Descartes' problem of dualism comes from his idea that *res extensa* and *res cogitans* are two separate substances, and that those two *cannot be known* in any other way but as separated (Descartes, The Principles of Philosophy 10-11) (Descartes, Meditations on First Philosophy 60-61).[1] How are we to reconcile such opposite claims of Descartes, and how can we understand him as a monist despite the prevailing and permanent problem of mind-body dualism in his philosophy?

If we observe what Descartes claims about God in *The Principles of Philosophy*, we could say that Descartes is, from the ontological or metaphysical point of view, a monist. This view is also supported by the claim from *Meditations* that mind and body, although *perceived* and *known* as distinct substances, are nevertheless in unity (Cottingham, The Mind-Body Relation 179-184); mind and body have to be in unity somehow, although we cannot know that unity *clara et distincta* (Cottingham, The Mind-Body Relation 183-184). God is also introduced in *Meditations* as a sort of verification of the existence of the external, sensible world, and not only as the possibility of the existence of beings, but as the possibility of knowing them (Descartes, Meditations on First Philosophy 25-37, 45-51). In various ways Descartes has attempted to present the possible explanations of the interaction between mind and body, trying to locate in human anatomy a possible meeting point of the

[1] Descartes was actually quite careful in using the term *substance*. It appears that the term *substance* is not very common in Descartes' writings, and when it is used by him it is used in a way that makes it clear to the reader that Descartes' concept of substance differs radically from the scholastic use of the term (Cottingham 65-70). Traditional Aristotelian and scholastic account of substance understands that concept in the context of the *substantia-accidentia* conceptual pair. If the mind is a substance, we can ask with Descartes, then, is thinking *an accident*? Descartes would firmly reject such a notion, claiming that *thinking* alone is exactly what *the mind as the thinking substance* is. In this article, however, I will employ the term *substance* for more technical reasons.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

two substances (*the pineal gland*) (Descartes, Strasti duše 182-183). Considering this, we can say that Descartes was convinced that mind and body are unified, are in some way intimately connected, as he himself states (Cottingham, The Mind-Body Relation 179), but failed nevertheless to give a clear philosophical account of that unity. What can we say, having in mind the above mentioned, about the problem of Cartesian *dualism*? What is there to be understood under that term?

The main philosophical questions in early modern philosophy are addressing the issues of the substance and the method. We may say that the question of the substance is *ontological* or *metaphysical* in nature, that is, it is concerned with *what there is*; the question of method addresses the issue of how to *know* that which *is*. The question of method, hence, may be described as *epistemological*, or more precisely *gnoseological*, in nature. Some philosophical traditions failed to recognize the difference between gnoseology and epistemology, claiming that they are the same philosophical inquiries, considering the same problem. Actually, in French and Anglo-Saxon philosophical traditions the conceptual difference between gnoseology and epistemology is absent (Filipović 117). Namely, both gnoseology and epistemology are considered as a philosophical account of knowledge.

Although it is true that epistemology and gnoseology are addressing the problem of knowledge, we may try here to elucidate the main cross point of gnoseology and epistemology: while epistemology asks what knowledge is, and what makes true knowledge different from false one, gnoseology questions how knowledge is possible and what the origin of knowledge is.

Epistemology would question what is already contained in knowledge we claim we have or could have, and how that differs from the content of other forms of knowledge, considering different subjects or ontological spheres; it *presupposes the possibility* of knowledge, that is, a *connection* between the knower and the object known, that way being closer to ontology. Consequently, epistemology questions the various types of knowledge, attempting to systemize them (Filipović 117, 378-379).

Gnoseology, on the other hand, and especially Descartes' theory of knowledge, which has its starting point in the *radical doubt* – not only in the possibility of knowledge of the object, but in the very existence of the object –, questions *the very connection* between the subject and the object of knowledge; it questions the very *possibility* of knowledge, its origin, value and limits, which is why gnoseological accounts are more un-ontological by nature, since they do not rely on *a priori* intelligibility of the object, but start from the knowing subject and its attributes alone.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, the aforementioned distinction between gnoseology and epistemology is, we might say, only an abstract and conceptual distinction; it seems that *in concreto*, in practice, no clear distinction exists, for every theory of knowledge shows a mixture of epistemological and gnoseological accounts. It is understandable also why this is the case; it is hard to conceive any knowledge without reference to some object, whether it is intelligible or sensual. In other words, it is hard to give an account of knowledge without being involved in or presupposing certain ontological framework.

In Descartes' case, however, there is a strong tension between metaphysical and gnoseological purports, for Descartes is pretty much convinced that the world exists in some other fashion, than it is perceived *clara et distincta*. For example, Descartes claims that mind and body must in some way be mutually interfering, although it is not clear how. Also, Descartes claims that *there must be some other* cause of the content of his consciousness, as well as the objects of sensual perception, although that cause is not known *clearly and distinctly*, for it is beyond cognition. God or the absolute substance is not described as a *concept* from which a metaphysical inference of the whole of reality can be made. It is *known* clearly and distinctly not God himself, but that *there must be God* (Descartes, Meditations on First Philosophy 26,30,32).

This Cartesian view shows the aforementioned tension between gnoseological and metaphysical, since the *concept* of *perfection* or *the perfect being* lies within the domain of subjective consciousness, thus implying a certain *gap* between *concept* and *existence*. If Descartes did hold that the concept of perfect being and perfect being are one, or, more precisely, that in God *his existence and his essence are inseparable*, then Descartes would have conducted his metaphysics and theory of knowledge from another starting point, but that would have led him nevertheless to dialectics. This kind of thinking is noted by Hegel, when finding fault with Kant's critique of the ontological proof for the existence of God (Hegel 92).

Descartes, however, similarly to Kant attempts to apply *the same logical categories to different beings of different ontological status*. Although Descartes reminds us that the term *substance* cannot be univocally ascribed to God and finite beings, he nevertheless holds that the only criteria of knowledge is a type of rationality similar to that of mathematics and geometry, where the concept must be conceived as *finite*, and thus *clearly and distinctly*. The absolute and perfect being, and thus unlimited in His power, cannot be conceived as a *finite concept*; it thus cannot be *defined*, and hence cannot be known *clara et distincta*. So, Descartes' theory of knowledge cannot be conceived as a starting point of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

his metaphysics, neither is it wholly compatible with it, if we understand metaphysics as an account of everything that is. His philosophy clearly shows certain incompatibility between the gnoseological parameters of true knowledge and his metaphysical claims. It is evidently a gap between his theory of knowledge and his conviction that God alone is the only absolute substance upon which every other substance depends.

If we could imagine a scenario where Descartes did attempt to conduct a metaphysical project on the basis of his theory of knowledge, then the concept of God, and mind-body unity cannot be included in that project. In that case every single segment of the metaphysical structure of reality must be compatible with the criterion of *clara et distincta*, and since mind and body are clearly and distinctly known *only* as separate substances, then the whole of reality would be conceived as a mystical interaction between two clearly distinct substances, or, in that case – metaphysical causes. Hence, Descartes' theory of knowledge hardly can be depicted as an epistemological project, as well as a metaphysical one. If we accept, however, that his theory of knowledge did hold certain metaphysical implications, those however would be insufficient to deliver a coherent metaphysical account, or to conduct a *mathesis universalis*, which Descartes did attempt. This is why Descartes' dualism was *a problem*, and not a solution, even for him.

In that manner, bearing the conceptual distinction between epistemology and gnoseology in mind, we might say that the dualism of Descartes' theory of knowledge is more gnoseological in nature, and it originates from the question of how to achieve knowledge, and whether knowledge is possible at all. Descartes' metaphysics and theory of knowledge are not compatible in a manner characteristic to, say, Plato, Aristotle or Plotinus, whose philosophies remain paradigmatic through the whole medieval period.

If we found it difficult to advocate that Descartes is, say, ontologically a monist, could we at least, while refraining from such a claim, replace it with another affirmative claim saying that – *gnoseologically* – Descartes *is* a dualist? The first claim is hard to defend, since ontology is not only concerned with *what there is*, but also with knowing what there is, and consequently cannot be divided from the method of knowing. On the other hand, Descartes never claimed that *everything that is consists of two principles*, namely, from the spiritual and from the material substances. Some authors do claim that Descartes was a metaphysical dualist, and that according to him, the world consists of material and mental substances (Heil 20). However, this interpretation of Descartes' dualism is wrong,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

since Descartes ascribes mental substance, thinking or *consciousness* to humans only.

It is evident from this that Descartes didn't hold that *res cogitans* is a metaphysical substance, but one from the gnoseological point of view. If *res cogitans* or *res extensa* is to be conceived as metaphysical substances, they would be conceived as whether absolute or finite; since God is the only absolute substance, then *res cogitans* or *extensa* can be conceived only as *finite* substances; and, as a finite substance, it cannot be conceived as the underlying or one of the underlying causes of everything that is. In the best case, it would be conceived as one of God's *attributes*, but in that case there would be no difference between Cartesian philosophy and Spinoza's, and then the whole Cartesian theory of knowledge should suffer a radical transformation. If we suppose that Descartes did ascribe mental substance to all objects of our experience, to all finite beings, then there would be no problem of mind-body dualism in the first place, and his philosophy would be a variation of panpsychism mixed with elements of Spinozism, as he would see matter and spirit as manifestations or attributes of one absolute substance, present in all beings. However, this is not the case. That way it is even more difficult to defend the claim that Descartes has advocated some form of *ontological* or *metaphysical* dualism on the basis of his theory of knowledge. Consequently, we can at least *via negativa* infer something about the monistic character of his ontological or metaphysical claims.

Considering all of the above mentioned we cannot but infer that *res extensa* and *res cogitans* can be seen as independent and distinct substances only in logical terms or in terms of *knowledge*, not in terms of *absolute existence as such*; nevertheless, Descartes held that *res cogitans may* as well exist without body the same way it exists with body (Descartes, A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences 29). However, since in terms of *existence* only God exists absolutely or independently, ultimately *res extensa* as well as *res cogitans* may at least have identical origin or cause, and thus *exists unified by that very cause*. So, when it is about mind-body dualism in Descartes' philosophy, we must refrain ourselves from inferring that we are dealing with some metaphysical system; rather, we are facing a gnoseological conundrum. This is supported by Descartes' claims in his *Discourse on the Method*, where he says:

> [E]xamining attentively what I was, I saw that I could pretend that I had no body and that there was no world or place for me to be in, but that I could not for all that pretend that I did not exist […] I thereby concluded that I was a *substance* whose *essence* or nature resides only in thinking, and

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

> which, in order to exist, has no need of place and is not dependent on any material thing. Accordingly this 'I', that is to say, the Soul by which I am what I am, is entirely distinct from the body and is even easier to know than the body; and would not stop being everything it is, even if the body were not to exist. (Descartes, A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences 29)

Here Descartes clearly presents from which perspective or ontological and epistemological modality he thinks; he does not claim that the body *doesn't exist*; neither has he claimed that, even though his or the essence of human beings in general is *thinking* alone, he *is* without body. The statement that body does not exist or that in humans body and soul *are* separated, and the statement that we have so clear insights and knowledge considering our "spiritual" nature without *any* interference of the concept of body are two substantially different types of statements, and those two reflect differences in both the metaphysical and the epistemological modality. Descartes is using words like "pretending" or "even if", which clearly reflect his gnoseological and metaphysical position; he does not say anything about the "absolute reality" of the object in question. In fact, "Descartes does not advocate the absolute reality of the content of consciousness *as such*, but only indubitability of their *existence*, that is to say, indubitability of the I as the general act of consciousness, which realizes those contents" (Petronijević 89; translation by A. R.).

Descartes claimed that mind and body are present in humans only (Kenny 212-216), but their unity is however hard to achieve through knowledge. Consequently, Descartes also does not advocate any form of *panpsychism*, claiming that in every material object there is also a soul or the *res cogitans* that gives that material object a shape, a form etc., which is more characteristic of older philosophical accounts (such as that of Ancient and Medieval philosophy), but he did claim that the only intelligible things in material objects are shape, their extending in space and motion (a claim that reflects a mathematical way of thinking, but also exhibits certain similarities to the (Neo)platonic and Aristotelian accounts of the relationship between form and matter) (Descartes, Meditations on First Philosophy 51). Maybe we are faced here with another form of dualism in Descartes' philosophy, namely, the dualism between the ontological and the gnoseological aspects of his philosophy? Obviously, there is a gap between Descartes' claims that everything that is originates from the one substance (God) (Descartes, The Principles of Philosophy 26) and the ones that stand for the mind-body dualism.

To present this problem in a clearer manner, we can try to clarify the way Descartes understands the *substance*. We might say that there are two

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

ways of Descartes' understanding of the term. On the one hand, the substance is that *which exists by itself*, and there is only one such substance, namely God (Descartes, The Principles of Philosophy 20). On the other hand, the substance is that *which is known by itself*, that is, that which does not require a notion of some other thing (Descartes, Meditations on First Philosophy 51-64) (Descartes, The Principles of Philosophy 20-23). The other is concerned with the mind and the body as distinct substances. Namely, in knowing the mind, there is no concept of matter/body involved. The same goes for knowing the body; in the concept of body as *res extensa*, there is no concept of the knowing subject and the mind involved. Those two are the only thing known as distinct from one another. Let us add to the list the third concept of substance, namely, that which considers *finite* beings; Descartes said that the term substance *cannot* be attributed to God and finite beings *in the same way*, but in doing so he *did not claim* that finite beings *are not* substances. So, the term *substance*, in Descartes, can refer to: 1) that which *exists independently*, 2) that which is *known independently* of the notion of something else, and 3) any finite being that *exists dependently* of the absolute substance.

We can see now that Descartes had no doubts that everything that is exists as connected in certain ways, given that everything that is originates from the one absolute being. Every finite being *exists* by means of dependence upon the absolute being or God. Despite this, Descartes did not found his theory of knowledge on those metaphysical claims, but rather on putting them away for a moment, while investigating the possibility of knowledge on the basis of the mind itself, or within the properties and limits of *lux rationalis*. So, having this "bracketing" and doubting as a starting point in investigating the possibility of knowledge, Descartes arrived at the brick wall of dualism between mind and body. How?

What seems problematic considering Descartes' dualism here is actually the way Descartes defines knowledge and the method of achieving it, relying on the criteria of *clara et distincta*. In his *Discourse on the Method* and in *Meditations*, Descartes gives a clear account of gnoseological and methodological criteria for considering some knowledge as *true*. We cannot speak of knowledge of something unless it is known *clearly* and *distinctly*; the fourfold methodological steps suggest that knowledge is achieved the moment the content of the concept is being sturdy analyzed in order to leave no hidden and unknowable leftovers (Petronijević, Od Zenona do Bergsona, Studije i članci iz istorije filozofije 207). So, the concept must be unveiled and evident completely, without any alien content that doesn't essentially belong to the object in question.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Thus clear concepts are *finite* and *distinct* concepts. The concepts known most clearly and distinctly are the concepts of mind as *res cogitans* and body as *res extensa* (Descartes, The Principles of Philosophy 10-11).

In the following section of this article I will try to present Descartes' gnoseological dualism as a consequence of the *reductive character of his theory of knowledge*, or gnoseology. I will try to present dualism, giving Descartes' philosophy as an example, as the consequence of the reductionism. This claim may seem odd, given that reductionism is an attempt to explain everything that is by means of reducing the multiplicity of the objects in question to a single explanatory principle. This reduction, however, is not without consequences and problematic issues, which I will try to describe in the following section.

3. Dualism as a Consequence of Reductionism

In this part of the article, I will consider the relation between the concepts of reductionism and dualism. The aim of this section is to show that reductionism is not opposed to the idea of dualism, or pluralism, but on the contrary, dualism may be considered as a result of reductionistic tendencies, as I will try to show using the example of Descartes' theory of knowledge.

As I have stated earlier, my intention is not directed towards rejection of reductionism, but rather on depicting its flaws and weaknesses, and before analyzing the nature of the connection between reductionism and dualism, I would like to be more precise on this point. By intention, reductionism is the concept that cannot be connected to the idea of dualism. It is an attempt to rather overcome any form of multiplicity. However, if we examine the very idea of *reduction*, we can observe that it logically refers to *negating* the subject matter, and then *translating it* using the other, more readable parameters. So it is actually the process of pilling of the unimportant layers of the subject matter, until it becomes a proper, readable and understandable subject. In other words, with *reducing* some object, the object is being *divided*, not *wholly* translated. So, logically, reduction does imply division; on the one hand, there is a set of properties of the object compatible with the parameters the reduction is being conducted upon; and on the other there is that *conceptual waist*, or incompatible properties of the object in question. From a historical point of view, we may also witness that reductionism has proved itself insufficient, or at least successful only for a certain period of time. If we are to create a certain standpoint, say, of *physicalism* or *naturalism*, we have to be very careful with the changes of the very concept of nature that

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

may occur, and the concept of nature, we may agree, is one of the most changeable concepts in the history of philosophy and science (Heisenberg 07-20). Virtually every concept is being subject to historical changes and shifts of hermeneutical horizons. Consequently, every concept that may serve at a certain moment as a point of reference for reduction is inevitably changing.

The concept of reductionism, or more precisely a *reductionistic tendency*, is a phenomenon appearing with modern philosophical/scientific ambitions. Maybe we could define reductionism as a sort of simplification or translation of one set of parameters using the other. That way, more complex theoretical systems or sets of values can be reduced, understood and hence explained on the basis of a simpler set of parameters, resulting in more "precise" knowledge of the object in question. In natural sciences, reductionism is evident in a form of *physicalism*, for example, which is an attempt to reduce all sciences to physics, including social sciences and humanities too (Kim 7269-7271). That would imply explaining social and psychological phenomena like existential crisis, exchanging of goods, communication, love, lust, hatred, and need in terms of laws of physics.

This might be appealing to some, since it encloses complex problems in a simpler manner. Also, we could infer that reductionism is nothing less than an answer to centuries old questions about the possibility of one all-encompassing theory, which, in early modern philosophy, was presented as *mathesis universalis*, science of sciences or universal knowledge. If reductionism is an attempt to explain various and complex theoretical systems using only one or as few systems as possible as its basis, is not reductionism then only a product of tendencies long present in philosophy to create an all-encompassing theory, since Plato and Aristotle? The answer to this might be, indeed, affirmative. However, in the case of reductionism, one must pay attention to certain moments that might be diametrically opposed to the very idea of *mathesis universalis* or universal knowledge.

What was characteristic for older ambitions towards achieving the "science of sciences", especially before early modern science, was not a reductionist, but a *holistic ambition*. Reductionism, however, is holism reversed! Holistic pretensions are probably as old as philosophy itself, or at least they appear the moment when man attempted to conduct a rational reconstruction of the whole of reality, which is known in the tradition of philosophy as *metaphysics*.

While holism, on the one hand, encompasses the multiplicity of elements trying to connect and explain them (Filipović 158), reductionism, on the other hand, attempts to reduce that multiplicity, and to simplify it.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, not only simplification of the multiplicity is present in reductionism, but *elimination* of that which appears as incompatible with *a priori* posed explanatory parameters. One early example of the consequences of reductionism in the history of philosophy could be located in Descartes' philosophy, namely, in the problem of mind-body dualism. Descartes, according to presupposed parameters of true and false, attempted to conduct the above mentioned rational reconstruction of reality, starting from the knowing subject and its faculties. That way, Descartes has to arrive at a dualistic standpoint, since an entirely new horizon of unexplained phenomena arose.

We have mentioned earlier that Descartes has attempted to apply the same logical categories to different beings of different ontological status; the criteria of *clara et distincta* may be valid when applied to certain concepts; most certainly, it is object similar to mathematical concepts or intelligible entities. However, when it comes to applying those criteria to the *unity* of mind and body, a problem emerges. Doesn't that suggest that the unity of mind and body differs ontologically and conceptually from mind and body alone? If the *clara et distincta* criterion doesn't appeal to that ontological sphere, doesn't that problem yield a different gnoseological approach or the resonance between conceptual and gnoseological? Descartes did not believe so, since he held that the only true conceptual approach is via *clara et distincta* criteria.

Descartes attempted to find an absolute standpoint for conducting the complete tree of knowledge, namely, a basis upon which every phenomenon that various sciences have as their subject-matter could be explained. This *fundamentum* Descartes saw in *res cogitans*, mind or the knowing subject. The absolute knowledge of the existence of the subject is present in the self-evidence of the knowing subject (Descartes, The Principles of Philosophy 11) (Descartes, Meditations on First Philosophy 17-24).

That knowledge is possessed as *clear* and *distinct*, and since it is a paradigm of true knowledge, the criteria of *clara et distincta* becomes an absolute criterion for everything that is to be considered as knowledge. That way, the criterion of true knowledge is posed by the knowing subject himself, and anything that bears an element of ambiguity or unclearness is dismissed as illusory or unknowable. When we talk about Descartes' theory of knowledge as *reductionistic*, we do not mean that Descartes attempted to *reduce* some substances to other (for example, to reduce body on the soul); what I mean by "reduction" is the *criteria* of knowledge, a certain type of rationality similar to that of mathematics and geometry, that appears at Descartes as paradigmatic rationality and type of knowledge.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

That which cannot be "scanned" with that type of rationality is not knowledge at all.

Although Descartes claimed that *res cogitans* and *res extensa* are separate substances in terms of knowing them, it is evident how even the concept of body is constructed according to the intelligible parameters posed by the *res cogitans*. The body cannot be known otherwise than that which extends, which can occupy a space. The extension or dimensions of the body can easily be calculated mathematically, which is why Descartes, in giving a scientific account of the body, uses the way of thinking or the same logic that is present in his analytical geometry. That way, the body is reduced to a set of mathematical parameters. This is the way matter or body can be known *clara et distincta*. In other words, *res extensa* can be known only as being filtered through intelligible parameters of *res cogitans*, or, more precisely, to the degree to which *res extensa* corresponds to Descartes' theory of knowledge.

Descartes did succeed to give an account of the concept of mind and body as distinct substances on the basis of the gnoseological criteria of *clara et distincta*. Where Descartes fails to apply these criteria is with the *interaction* between those two distinct substances, since it is evident that those concepts exclude one another (Descartes, Meditations on First Philosophy 51-64).

This demonstrates that these criteria cannot be universal, since there appears to be a set of phenomena of our experience – like the one that refers to mind-body connection; for instance, a sensation of pain, or hunger (Descartes, Meditations on First Philosophy 53) – that cannot be explained by these criteria. Instead of abandoning the initial gnoseological project of formulating the criteria of true and false, Descartes continued through his life to pursue the question of the mind-body unity based on the aforementioned criteria.

This is an example of how reductionism multiplies its work; instead of giving an all explanatory account, it creates a whole new set of hard-to-explain problems, which is why it causes new dualistic or even pluralistic positions. Reductionism is indeed an attempt to overcome any form of dualism/pluralism. However, what I try to point out here is *the final consequence of reductionism*. Regardless of intent to eliminate any form of multiplicity, reductionism prepares a new basis for another form of dualism. The reason is that reductionism is not holistic, but exclusive; namely, it excludes and eliminates those phenomena that are hard to explain on the basis of criteria upon which reduction is being conducted. Those neglected sets of phenomena are the ones that cannot be reduced,

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

and hence explained, the same way the mind-body unity in Descartes cannot be reduced to the criteria of *clara et distincta*.

4. Concluding Remarks

The example of Descartes' theory of knowledge demonstrates the way reductionism, as an attempt to overcome any form of multiplicity and ambiguity, actually makes another form of dualism, and poses other ambiguous problems. Descartes was a dualist in the sense that he was unable to explain a certain set of phenomena of our experience relying on parameters of his theory of knowledge only. He does not, however, advocate any form of metaphysical dualism, and hence mind and body are not to be understood as metaphysical principles or causes which everything that exists consists of.

Not only Descartes' reductionism causes 1) gnoseological dualism between mind and body, but also dualisms within other spheres of Descartes' philosophy as well; for example: 2) a dualism of monistic metaphysical (God is the only absolute and independent substance upon which every finite being depends) and dualistic gnoseological claims (mind and body are two separate and independently known substances, whose interaction remains unclear), or 3) dualism between sets of phenomena of our experience *explainable* with gnoseological criteria of *clara et distincta* (*res extensa*, *res cogitans*, or mathematical parameters), and sets of phenomena *unexplainable* according to aforementioned criteria (such as the mind-body union, and hence any other phenomenon related to that union).

Bibliography

Cottingham, John. *Cartesian Reflections*. New York: Oxford University Press, 2008.

—. "The Mind-Body Relation." *The Blackwell Guide to Descartes' Meditations* (2006): 170-192.

Descartes, Rene. *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Trans. Ian Maclean. New York: Oxford University Perss, 2006.

—. *Meditations on First Philosophy*. Trans. Michael Moriarty. New York: Oxford University Press, 2008.

Descartes, Rene. "Strasti duše." Kangrga, Milan. *Racionalistička filozofija*. Ed. Vladimir Filipović. Trans. Milan Kangrga. Zagreb: Nakladni Zavod Matice Hrvatske, 1979. 178-204.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

—. *The Principles of Philosophy*. Trans. John Veitch. Munich, 1901.
Filipović, Vladimir. *Filozofijski riječnik*. Ed. Vladimir Filipović. Zagreb: Nakladni Zavod Matice Hrvatske, 1965.
Hegel, G.V.F. *Nauka logike I*. Trans. Nikola Popović. Vol. I. Beograd: Beogradski Izdavačko-Grafički Zavod, 1976.
Heil, John. *Philosophy of Mind*. London and New York: Routledge, 1998.
Heisenberg, Werner. *The Physicist's Conception of Nature*. Trans. Arnold J. Pomerans. London: Hutchinson & Co.LTD, 1958.
Kenny, Anthony. *The Rise of Modern Philosophy*. New York: Oxford University Press, 2006.
Kim, Jaegwon. "Reductionism, problems of." *Routledge Encyclopedia of Philosophy* (1998): 7269-7271.
Lee, Richard A. Jr. "The Scholastic Resources for Descartes's Concept of God as Causa Sui." *Oxford Studies in Early Modern Philosophy III* (2006): 91-118.
Petronijević, Branislav. *Istorija novije filozofije*. Beograd: Zavod za udžbenike i nastavna sredstva, 1998.
—. *Od Zenona do Bergsona, Studije i članci iz istorije filozofije*. Ed. Slobodan i Marić, Ilija Žunjić. Beograd: Zavod za udbženike i nastavna sredstva, 1998.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER SIXTEEN

HOW HUMAN BEINGS WORK…

JAIME MILHEIRO

(TRANS. DIANA NEIVA)

"I think… therefore I exist… therefore I function…

… I function because I feel myself functioning… because I am conscious of it… even when I equally feel that much of what is happening with me escapes me (unconscious)."

Like all living beings, human beings "know" they exist. They feel it, with a marked difference relatively to the others: they include emotional and affective roots in that knowledge. Roots of various perfumes, which absolutely integrate them, without any exit portals. They simultaneously recognize that everything they process in that intimate functionality, besides the moment's performance, has only one end: to promote well-being and extend life time, if possible with pleasure, never being able to escape or ignore it.

This means that, as a result of the reached evolution stage, each human being disposes of a very personal subjectivity and an inescapable ability to feel and to look at themselves, besides of a particular ability to look at the Universe. In other words: he is a person… he disposes of a "psychic apparatus" where his look, his knowledge and his thought develop… an "organ" he slowly formatted during his childhood, and where, not knowing how to read or write, he structured so complex sensibilities and so global operations that only in theoretical and poetic exercises is it possible to fragment the set and separate it from the body.

In the concepts of "Structural Psychosomatics", "Mysteriousness", "Meaning of Path" that I have been trying to conceptualize, I stress how the historical "invention of the soul" tries to deceive that overall exercise and promotes absurd separations. I remember the damages such an invention brought to people, justifying wars and making them "holy",

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

through an ineffable promise of a journey to God's side, the only One. The salvation of the soul will be worth anything in such a conception, including the exclusion and death of others, indifferent to those details. An example at hand is the so called Islamic State, whose practices are well known, and that only exists because the soul of its believers will fly to heaven after they die. Without that flying soul, identical to the soul of other religions, they would not kill nor die, although strangely our culture excuses itself from arguing about it. It only pretends, in the tacit acceptance of who is identically compromised, having no interest in critical or pedagogical movements. Their interests are other ones. They are from a different "religion".

Commenting the role of the brain and computers in such questions, I remind that the ability to think belongs to the person, not to their brain, even if it is absolutely indispensable in its supporting processes. To deepen the knowledge of the brain thinking that doing so the knowledge of the person is deepened would be the same as studying the ear or its auditory functions thinking of that as a way of getting to know the subject's music and musicality.

I also remind here that a computer does not make mistakes, it does not judge, nor does it have the unconscious. Therefore, it can never think beyond what the human put there, since the thinking mind (thought generator) is the result of the inter-conjunction of history, memory and emotions involved into the functionality of each one of them, and of the insertion of the respective contents in the data granted by evolution. Not in machinery stripped of feelings of the past, as sophisticated as they might be.

In their vital circuit, human beings bound themselves to a progressive scale of necessities. The more basic one (survival) is followed by many others, more and more elaborated. Needing an identity, affection, connection, safety, trust, fulfillment, etc., everyone moves in that direction. But because they are prematurely born and totally dependent on who warms them, in the first years of life they only survive in a protective relation with someone, bounding themselves, since they are infants, to internalize the vicissitudes of all they lived and suffered in their growing and development process.

Mother, father, family, teachers, idols and other significant adults turn out to be references and objectives of each one's own identification. Each child assimilates the good and bad facets of the relationships he had with them, of their models and principles, of their gratifications and frustrations, constructing little by little a building dependent on that and very hardly changeable.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

It will be in the subjectivity of these joys, sufferings and anxieties that their future readings and operabilities will be based on. Their level of reality, fantasy and freedom will guide the person through their life by mechanisms that, being partly rational and conscious, will equally contain unconscious processes, given the fact that it is impossible to live and be conscious of everything at all moments.

There are fundamental features of the species that work in this organization and that end up assuming decisive importance to the emotional perfume of each one of us.

The ability to mentally represent an object, to "see" it with the eyes closed, the ability to symbolize, to postpone the desire, the ability to distinguish an affective relationship from a destructive one, the ability to make considerations about the beautiful and the ugly, the ability to make mistakes, judge, fantasize…

… these are abilities easily recognized in all human beings…

To those I add another one we don't usually talk about, which is "mysteriousness"… a certain magical charm about the unknown without which no one could live… the one that "religiousness" tends to sacralize and that "religion" has historically taken as its own, closing it in dogmatic frames.

Such mysteriousness, a mix of fear, perplexity and curiosity, is, to me, one of the fundamental features of the species. Being an inevitable (favorable) result of the fear of the unknown that always exists in the early relationship with the mother (the fear of losing her, of dying before the threats of the "uncanny"), it is the most intimate propellant of what distinguishes us from the savannah cousins, only instinctively moved by hunger and the estrus cycle. It is what gives us the pleasure of functioning, not only the obligation of functioning before necessities.

Counteracting the fear of the unknown, mysteriousness gives us approximations instead of exclusions. It approximates us to the "other", different from the mother, for reasons that have nothing to do with sexuality or aggressiveness, turning it into a desired object of pursuit, of knowledge, of hope and creativity. Its exercise solves countless sufferings, although it can also flow into beliefs and disquiet that blur the subject, detracting clarity from him.

Without it, the other would not be more than an organic thickness to fear and exclude.

By means of idealization, one part of that mysteriousness is sacralized and is transformed into "religiosity", leading humans to longitudes of eternal contemplation and to "other world mysteries", normally in such insoluble and ruminative terms that they never end. Mysteries that when

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

cataloged in dogmas don't even admit pauses, under penalty of divine punishments.

Homes of gods, Olympuses, paradises and other fantasies countervailing the fear of death multiply anxious destinies, fomenting organizations intended to do that, and brotherhoods so avid of themselves (religions) that antagonize each other fiercely, given the fact that they depend specifically on "revelations" of their God, the only one among the only ones, as is highlighted above; all of them preaching, moreover, concepts of a "spirituality" reserved only to its "faith" and to its "transcendence", as if what we all do – thinking, dreaming, imagining – was not a complete spiritual action disconnected from such parameters.

Nobody is born with or without that, such as no one is born with health nor disease. The baby carries formless elements on which he will elaborate, in the initial relations he absolutely depends on, his building abilities.

Through his journey he progressively considers ethical and moral purposes, still with no name, that only in very disruptive circumstances don't emerge, building a first court inside of him. An internal court, way prior to civil and religious courts that he will later be confronted with, conceived only for the solving insufficiencies of the former.

The judgements and "blames" of that first court will be very important in the functionality of each one and in their practical movements of life: very decisive in health, disease and behavior. An eternal scribbler of balance sheets about the done and the undone, about himself and the others; through him all humans generate idealizations about this irremediable paradox evolution created to them: being capable of thinking, desiring and promoting much higher than what they are effectively capable of reaching.

Inventing galaxies for such accesses, galaxies no one knows where, humans then mystify their own scopes and their own culpabilities, packing mysteriousness into indemonstrable positions. They divide body and soul in two, and make culture about it, even if they subject themselves to the demands of deities they create and, in their view, compensate and recompense them.

A good mental health is probably the organized functioning of all that texture in each one. It probably is a feeling of an active and positive well-being in the current flow which is only noticed when it fails; a feeling which, in the interaction with others, expands and atrophies, gratifies or complicates, depending on the feelings of internal freedom each one has.

With more or less cohesion, it probably is the summary and the global experience of a triple capacity: the capacity to relate to oneself, the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

capacity to relate to others, and the capacity to effectively live the circumstances in a process initiated in the childhood and with a permanent dynamic.

Always aware, the lack of only one of these parcels is enough to raise doubts, blossom anxieties, and make decompensation happen.

A growing child goes up stairs: gradually he begins to be able to look into the known and the unknown, to think about himself and about others, to process desires, joys and sorrows, at a never fully satisfactory way. He deduces mathematics and philosophy without knowing how to say them, he links horizons which he's beginning to realize to be endless. All this provides him with excellent conditions to continue in mysteriousness and to invent in dissatisfaction. He will tend to fix what he aims in mysteriousness and does not find in life, through a fermentative imagination which easily admits the infinites that religious formulations add in the format that suits them.

In the course, directed right or wrong, this child settles three psychological conditions: feelings of autonomy, feelings of responsibility and self-feelings. It would not be indispensable, but by norm he regiments such conditions to the sacralities that meanwhile he has been provided with against glimpsed anxieties. It has been like this historically and it will continue to be like that, until we can suppose some more clear solutions.

From that to saying he has a “soul” separable from the body is a little step – it is convenient, affirms it and compensates it. Attributing displacements to less anxious and more recommendable places to this same soul will be another little step astutely cultivated by the “religious philosophies of mind” which attract such purposes with dazzling resurrections.

Other acquisitions, such as the decision-making and ability to chose, a sense of adequacy and justice, respect for others, tolerance to frustration, acceptance of the difference, absence of victimization, ability to be alone, will root in these pillars, shaping a set that will preside over all his balances and imbalances.

Always unstable, more fragile or more resilient according to the cohesion and coherence acquired, balance will also depend on the invasive dimension of the circumstances and the resonance provoked by them, so Mental Health will never be a building whose virtues and vices could be presented in geometries and numbers, like selling or buying laboratory chemicals.

A “Path Sense” is probably the realization of all that, in an integration of the internal time into the time that passes. It is to feel identical to oneself, in the continuum of past, present and future, being certain that one day it will end.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

It is inside that such feeling circulates. It is by making their path that the child builds it, in a healthy format, in an extension of what he feels and in a process which includes the idea of death and respective anxieties, that the more disguised they are, the worst: they will make the child more vulnerable.

Masks and absences engineer interiorities in which words lose sense and shorten purposes, maybe they are even entangled in glorious expectations intending to forget.

Given all this, the human being can only be a "Psychosomatic Workshop" of permanent laboring, that constantly seeks to reduce suffering and build Health. All his emotions, affections and feelings, as all his biological and physiological features participate in this process, trying to avoid discomfort and fortify an identity that will always be a "Psychosomatic Identity".

The concepts of "Structural Psychosomatic" and "Psychosomatic Fact" about which I have been theorizing dismiss the dualism above referred, seeking to know better what we are and how we function.

The sciences that instructed us and that we are part of have been clarifying us in many of these questions, but they also have inherently limited us. Such limitation transpires in the difficulties that we always have when we want to assume a unitary functioning. Inside us, as within all medical science and all psychological science, there is a strange resistance in this regard, derived from the dualist account that shaped us. We even very hardly arrange instruments or words to help us.

For example, all the concepts of interaction, reciprocity, influence, equivalence, homology, simultaneity, interrelation, conjugation, representation, etc., often used when one wants to theorize about what happens inside the body and the mind, they do nothing more than maintaining, in a latent way, the historical readings in which dualism persists, as if we were irredeemably conditioned by it.

Gathered in this medieval separation body/spirit, we have partial answers. But, if they are not presumably random in the functioning of the human being, why are the consequences of psychological suffering so different in one's body from another's?

Why, for example, accumulated angers are accompanied by skin disorders in some cases, by heart disorders in others, by depressive manifestations in many others… and in others nothing is triggered?

Why, in situations of loss, are there people suffering from headache, others from digestive disorders, and many others from nothing at all?

Why will one wake up extremely weak in the muscles and be like that for a few days without even having moved in his bed, one who made such

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

a physical effort dreaming that exceeded all his limits?

Why will one ejaculate with an orgasm when he dreams, without any bodily movement made?

Why, in so called "fibromyalgia", people with a sedentary lifestyle who in their mind make intense runs have body aches?

How is this all guided… how and why?

We don't know very much about it. And it won't be the current neurosciences, nor a few superficialities so called neuropsychological that will help us to discover the "psychosomatic markers" that such integration presupposes. Only a new scientific paradigm can guide such paths, in my opinion.

Probably, the future will consist in the investigation of an idea as simple as this: the dynamic process of identification, absolutely essential for the growth and characterization of the human being, probably does not consist only of the psychological internalization of what the child admires in the other and wants for himself, as it is said. He probably also carries markings on the body. Markings of other nature, markings of "corporatization" (incarnation), markings of which there will be structural signs still undiscovered.

If this corporatization (a very different concept from the psychological concepts of "incorporation", "internalization" or "introjection" which refer to an imaginary body) does not meet the desired flow, the early sufferings will not elaborate enough and will propitiate maladjustments that will turn into disease. In my conception, the disease will be located in the area of the body where such maladjustment has been triggered, therefore always being a psychosomatic manifestation. Separating body from spirit is as absurd an irrationality as separating hydrogen from oxygen in water. It will be impossible to do so.

Nowadays it is said (they try to say…) that the brain is the only defendant of the whole process. Intoxicated by the superpower we attach to it, we deify this brain, although historically we have heard worse things. In other times, the defendants were God, the Devil, the soul, the unconscious, the genome, the DNA, and so on. The human being has such a need to find the guilty of everything that happens to him that, as the gyrus are exhausted, he will certainly tend to find others in the quantum physics of his scientificity and in the laborious temptation of his non-responsibility.

Being healthy is to sail without grumbling and without blaming, as the stream of blood that flows in us. It is my own subjectivity, built by my story, flooded by my memory, that qualifies and quantifies me. It is its weirdness, so inappropriate and so little cooperating that does not even let

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

itself be configured by the most recent and expensive technologies, that gives me the right to exist. It is not my brain that instructs me to exist. It is, though, my subjectivity that makes me feel here, functioning, knowing me here, identically to all of you.

At this point in which, as it is said, the human being is nothing more than a neurologic machine of walking balances and fragile connections, having the quantitative of freedom and lucidity that we had achieved through centuries removed, it even seems that they want to take the ability to feel as individuals who are able to elaborate from us.

It seems they want to deny our existence, that is, our feeling of functioning as a person. But only those who want to will accept such an abuse, stated as scientific. Obviously.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Chapter Seventeen

The Human Being and the Ancient Philosophies of India

José Antunes

(trans. Diana Neiva)

Addressing the history of philosophical and religious thought in India we find ourselves with a diversity and originality inherent of a territory so vast and multifaceted that these landscapes seem to influence the way of comprehending the cosmos and the human being. Since the ancient times of the Vedic period to the course of the first millennium A.D., systems of thought emerge always seeking to solve the old riddle of the Sphinx: who am I?

Founded on the Vedas, all these ways of understanding the nature and humans show a creative richness of fertile imagistic resources, always with a desire for a mystical union with the Unknowable, with a capacity to open up to new ways of seeing the world, expressed in "heresies" that were also born in these lands, such as Jainism and Buddhism.

Indian classical literature exceeds in production what was produced in the classical West. The two famous epics, Mahābhārata and Rāmāyaṇa, are in extension about fourteen times the Iliad and the Odyssey together! The sayings or philosophical precepts synthesized in the Upaniṣads or the Yogasūtras of Patañjali provide numerous pages of profound philosophical reflections...

The schools known as *darśana* cover the various possibilities of knowledge from a more spiritual vision, the *puruṣíaca*, to more material atomistic views, *prakríticas*, in a Democritus way. Thus, views or perspectives about Nature and the Whole unfold, which try, sometimes with panting eagerness to synthesize mystics and thought, to give light to the inexplicable of human nature and essence.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Supported by the philosophy of the Vedānta, which only admits the Supreme Spirit, Parabrahman, as Reality, everything else being mere illusion of our senses, we will make an explanation on the nature of the human being with his various constituent parts. From the unity of the Being, to mind-body duality or body-soul-spirit trinity, we will observe the seven parts that constitute the complete human being, according to this philosophy. As layers that cover us, these dimensions which exist in increasingly subtle vibration levels, as we go from the material to the spiritual, are a chance to understand this puzzle we are in more depth.

Assuming that the only reality is that Supreme Spirit, the denial of everything else that is manifested is not done in a nihilistic way, but by considering the manifestation as something transient, thus changeable and subject to illusion. A plural and illusory world, intensely lived in every moment of space-time as a fleeting reality. It is sought that consciousness progresses from the experience of the temporal to a more and more lasting reality, from the experience of the effects to the increasingly perennial causes that generate such effects, understood as Laws.

Man is a part of the Being; he is a manifestation of that ulterior Essence we call Being. A Metaphysical Reality, often far from a mental conception but inevitable for the construction of our understanding of the universe. A mystery behind all the visible, but also a Theos behind the Cosmos, which can be captured through the Logos, the language, the understanding, the Reason.

The Being shows up through the universe, it appears... and the Existence presupposes an Essence. And so it is in the case of the human being: through existence, we can go up the stairs of understanding until we are closer to the essence. Closer to us, or more easily visible to our understanding, are the existential aspects, and by their observation and study we can better understand what we are.

The human being is not disconnected or separated from nature, and thus the principles and laws that lie in it must be present in man too. By the observation of nature, various levels or classes of manifestation were found, that the Hindu sages, but not only them, synthesized in seven, which, in a simple empirical observation, we can see in the expression of colors and musical notes. As if nature were made with a number basis founded on the number seven, which could justify the "mystique" weight that some attribute to this number.

Man is a synthesis of nature, he's a microcosm, that is, the laws of the macrocosm are reflected on him. Was the old Greek aphorism inscribed in the Delphic temple correct: "Know thyself and thou shall know the Universe and its laws"? Is the human being the key to the solution of the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

puzzle? We cannot prove by our own experiences what we haven't lived yet, however we can admit that others can have experiences of realities distant from us. And the possibility to experience is always within our reach if we make every effort to that.

Let us be guided then by the Vedānta philosophy in its explanation of what a human being is or, more specifically, of the forms or "garments" that the human being uses for his existence.

Starting by the "foundation" or basis, we find the physical, dense, material body which was called sthūla-śarīra in India, whereby śarīra means *vehicle*, instrument, possibility, and sthūla dense or that which can be grasped. It should be taken into consideration that this instrument is not only physical matter, the accumulation of atoms that build our physical body, but also that "electricity" or magnetism that enables the aggregation of these atoms and the maintenance of our physical appearance throughout life. It has the characteristic of being tough, hard, and because of this it was associated with the element Earth and also with the mineral kingdom, which, with their ability to withstand other elements of nature, simultaneously and apparently, convey the image of the unmoving eternity.

But this heavy body needs a breath, an energy that makes it move, grow. This possibility we have is called prāṇa-śarīra, a body or energy vehicle related to the element Water. It contains the power of development of the physical body, for, without the energy, it would be like an inert and lifeless body. The symbolic relation with water is very direct, as it is always life, always movement, it is an ocean as a primordial *mater*, and rivers as fertilizing channels of the land through which they flow. It is the source that moves all that is visible... and the plant kingdom represents well this energy dimension, always eternal, with this possibility of multifaceted growth.

But here another dimension comes into play. In our scenario we have so far a world of matter as a basis, support, upadhi in Sanskrit, of a life that is manifested by the growth of forms; but now comes the plan of movement, of motivation. In India it was called liṅga-śarīra, vehicle of the emotions associated with the animal kingdom. Now it is not enough to be rooted to the earth, now the scope of the motivations that lead beings to approach something comes into play, what they need for their subsistence and their realization. In humans, this dimension allows emotions and feelings, the motivational movement that leads us in pursuit of knowledge of the things of the world to achieve self-fulfilment. Its symbolic relationship with the element Air leads us to realize the natural instability of our emotions, this game of "I like" and "I don't like" and "I like again" we are subjected to, and that, not being false, we reflect this instability of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

the "atmosphere" always influenced by the winds that blow. It is important for humans to the extent that it is the foundation of their motivation through the path of life, with its always dangerous limit states of euphoria and depression.

The fourth vehicle is already specific to human nature for it is constituted by the mind. In the East it was called kāma-manas, mind-desire, and it is a concrete and objective mind, that is, directed to the external world. It allows us to organize our everyday lives with their basic concerns with our survival and passing through the world: feeding, job, socializing, fun, etc. They distinguished this feature as typically human above other animals' capabilities and then they related it with the element Fire: Man is the only being who masters and works with fire, and this, symbolically, represents this mental dimension so distinctive of mankind.

With these four components of the human being, the physical body, the energetic, the emotional and the mental, they claimed that the basis of the manifestation of Man was built. Closed in four elements, crossed in an existential manifestation of space-time, he was fit for the conduct of the drama of life. However, these four vehicles are subject to the laws that govern the levels of existence and are thus subjected to time, that wears everything out. They constitute the personality that changes and improves through the path of life.

But above this concrete mind, the kāma-manas, that all human beings possess with more or less acuity, they asserted we can access through our own effort another instrument that exists in us: manas, a pure mind disconnected from personal interests that usually are the field of action of the previous one. A more abstract mind with the capacity to reflect about objects without taking into account the mere personal viewpoint. A mind that elevates itself and seeks to see with a more extended scope. Whereas the “sentry post” of the kāma-manas, its basis for observation/reflection, is its personality, i.e., it is influenced by the egocentrism inherent to personality, this manas seeks to elevate the basis for observation, seeks to rise ourselves so its horizon turns into a more vast one. Many times we feel as a fact that philosophical reflection is not easy, or rather difficult when something affects the physical or emotional body, and this exercise of elevation of conscious is the foundation of philosophy.

This mental capability, or what we call mind with its physical "tuner" which is the brain, is characterized by functioning or developing on the field of duality. It is impossible to conceive of the day without bearing its opposite, the night, in mind; the same occurring with the high and the low, or hot and cold. For the Pythagoreans the mind was governed by the number two and worked with truth and falsehood to make a path of choice

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

relegating what's less true. Or, otherwise, this duality manifests itself always in the presence of a subject who observes and an object which is observed. For Vedāntins, Man could access another vehicle above this dual mind: they called it buddhi, lighting, which we can associate with what we know as intuition. It is Archimedes' eureka, the "I got it" when we can solve a problem or difficult puzzle. As a vehicle more in potency than in act, it is difficult for us to find it in us and observe it as we observe and analyze the physical body or the emotions. The Pythagoreans associate this form of knowledge to the number one, a direct knowledge of reality without the intervention of the analyst mind, a form of union with the being of the observed object.

Many times the advances of mankind were achieved by scientists, artists, inspired human beings who "captured" something and then, by a mental process, transmitted that to others.

Lastly, the Vedānta philosophy tells us about a mysterious feature that is reflected in the human being: ātman. We can translate this word as spirit, however, this term is very broad and not easily definable. In fact,"spirit" might often be confused with vibration levels, material too, but more subtle, less dense. In this respect Vedāntins are clear and assertive: all the Reality that matters is the great mystery, Parabrahman; the rest are the appearances that this Reality wears. In this perspective, the vibration levels of nature less dense than matter, therefore less observable by science, the same levels that appear in Man as energy, feelings, thoughts, etc., would be, according to the Vedāntins, no more than more subtle aspects of matter, but not the spirit. Something deeper is then manifested in us, and they said that ātman was our essence, the presence of the Being in us, our true center or what hides behind the Self…

Making this vehicle, power, feature, etc., which is present in every human being as it is its essence, present, that is, passing what is in potency to act, would be a task to everyone who wants to perfect and apply every effort to achieve this. They said it isn't easy, on the contrary, and the concept that wisdom is for the elected ones has its foundation here, not to the extent that they were elected by "someone" or for some special condition at birth (basis of all racisms), but because there are few people who want to develop by effort reaching the possibilities nature has put inside of us. At all times, the mythology of heroes has its root here: heroes or demigods who had self-knowledge and conquered themselves experiencing all that Nature has placed in us.

But let's go back to ātman to note that, although it is the last aspect inhabiting the human being, at a high and difficult to reach top, this power may be expressed or present itself in the form of will. This will is a feeling

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

whose origin is unknown, but we know it *is*. It should not be confused with a mere desire that drives us to something and crumbles at the first obstacle, it is instead an inner impulse able to make a path overcoming obstacles, a will that moves mountains, a will that is the motor of a greater motivation not limited by circumstances. Even more, the sometimes difficult circumstances are opportunities to demonstrate if the motivation is real or not.

Human beings are constituted with these seven characteristics, supports or vehicles. If they are fully functioning, then we have a perfect human, or at least closer to what we conceive of as perfection. Nothing is static, everything is transformed, immutability is characteristic only of Parabrahman; and in this movement which is Life we move into a direction, in a sense, the Cosmos is governed by dharma, and the human being participates in that movement. Self-awareness and self-control are the best marks for us to realize this directionality.

In the history of Indian thought currents, despite the plurality of views, often opposite, we find this concern to go further, a dissatisfaction with what we have or what we are, seeking to apply effort to overcome it. This boost for achievement and union with what we lack, this attraction for what we don't know and what we don't have, may be the origin of the mystic impulse so characteristic of these eastern lands. Maybe it's also the source of attraction or fascination that the East has always provoked in the West, be it today in a search for mystique, or at the time of the Discoveries in the search of Prester John’s Kingdom.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER EIGHTEEN

THE FUTURE AUTOMATION OF THE BRAIN: A NINETEENTH-CENTURY POLEMIC ON THE PERFECTION OF CONSCIOUSNESS

MANUEL CURADO

The present chapter studies a controversy that happened between two medical doctors in the late nineteenth century. Dr. José de Lacerda (1861-1911) advocated in various publications a theory of the conscious mind that is worthy of reflection. Consciousness emerged early in the evolutionary process but will disappear as evolution progresses. The future of evolution will have beings devoid of consciousness and automatic brains. Looking at the present and past of evolution, the existence of consciousness is considered an imperfection. In the preface to the work *The Neurasthenics* (1895), Sousa Martins (1843-1897) surpasses what is recommended by the protocol of prefaces and criticizes the conception of the evanescent consciousness of the future. The chapter analyzes this intellectual debate and emphasizes its interest to the philosophy of mind. It is argued that a future without consciousness is a manifestation of the desire to live a present without consciousness, avoiding what Lacerda calls neurasthenia and evil of living (in French, *le mal de vivre*), and what could be called the fear of being conscious.

I

At the end of the 19th century a curious controversy took place in Lisbon about the role that consciousness has played in the evolution of biological species and what it may have in the future. The physician José de Lacerda (1861-1911) defended in several publications a theory of the conscious mind that deserves reflection. From the point of view of this Azorean medical doctor, consciousness arose early in the evolutionary process but will disappear as evolution progresses. The future of evolution

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

will have beings devoid of consciousness and with automatic brains. Looking at the present and the past of evolution, the existence of consciousness is considered an imperfection.

Lacerda seems intellectually fascinated by the idea of a robotic humanity. His idea of human perfection is that of people who behave exclusively unconsciously, without feeling anything. For Lacerda, when a new task begins, consciousness becomes especially intense; when the task is mastered, consciousness ceases to exist and the task can be performed automatically. This is the perfect model of the explanation of the relations between unconscious and conscious mind in the human species: "The consciousness, maximally alive at the beginning of the respective neuronal learning, becomes more and more blunted the more perfect it is" (1897, p. 176). Consciousness was present at the beginning of life on Earth, but with evolution and improvement over millions of years, some beings have managed not to be aware, and the future will be precisely the generalization of these success stories. The perfection of organisms will happen in the future when everyone becomes unconscious. In his own words, "consciousness is a childish and decreasing way of neuronal sensitivities ... it is an inferior and transitory state of nervous receptivity, indispensable for the attainment of automatism" (1895, 27, 1897, pp. 124-125). This surprising theory runs counter to the usual perspective of the distribution of consciousness in the world of life, that consciousness seems to be linked to higher animals and humans. Less than a decade before Lacerda, the American philosopher William James described differently the relationship between consciousness and the various levels of complexity of biological beings: "It is very generally admitted, though the point would be hard to prove, that the consciousness grows the more complex and intense the higher we rise in the animal kingdom. That of a man must exceed that of an oyster" (1950, p. 138). It is idle to inventory the difficulties James equates with the literary elegance that characterizes his prose: the distribution of consciousness in the world of living beings; the proof that all living beings are sentient; the proof that the phylogenetic past of all living beings has been uninterruptedly characterized by consciousness; the knowledge of the mind of other beings (in the sense of access to their subjective point of view); the impression that consciousness accompanies the level of complexity of beings, that is, that it is more intense and rich in beings more developed than in the simpler beings; the problem of taking the characteristics of the world in a certain period (the present time) to formulate conjectures about the characteristics of the world in other historical periods (the past and the future); etc. Lacerda's conjecture that consciousness is tied to primitive beings early in life on

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Earth is dissonant with what is usually considered self-evident, as seen by this thoughtful view of James. This conjecture is based on the interpretation of the previous sequences of the biological evolution and is complemented with a preview on the future of the consciousness of the living organisms. In the first place there is the idea that living without feeling anything is an advantage over living with awareness of something. In the second place there is the conjecture that the human brain will no longer be conscious in the future; and this is interpreted as a sign of perfection of this organ and, of course, of improvement of human life. Reflection on the way in which individual human beings learn new tasks may have influenced the conjectures that have been proposed both on the past and the future of evolution and on the alleged advantages of living without consciousness. In this case, one could say that it is from the phenomenology of the human being that Lacerda explores the problem of the presence of consciousness in various periods of the evolutionary history of living beings. It is evident that the problems James recognizes are not answered by Lacerda; however, it is hard to imagine that one can do better than him, that is, to take on the characteristics of how the consciousness of human beings at the present time is to conjecture how it could have been in the distant past or how it might be in the distant future.

This brother of the musician Francisco de Lacerda (1869-1934) was an unusually bright physician who completed his clinical activity with activities in other areas, such as the writing of poetry books (e.g. Lacerda, 1891) and studies on the cultural life of the end of the nineteenth century (Lacerda, 1901). In medical context, he was one of the first clinicians to criticize the domain of the degeneracy theory in finissecular psychiatry. Although his life has terminated early due to tuberculosis, the foundation of the theory of future brain automatism is an important part of his intellectual work. In several publications he presented arguments to justify his conjectures, namely in the book *The Neurasthenics* (1895), dedicated to the reception of the ideas of the American neurologist George Miller Beard (1839-1883) on neurasthenia, published only one year after obtaining the degree of Medicine at the Medical-Surgical School of Lisbon; in an article entitled "Hypnology" which was published in 1897 in the journal *Arquivos de Medicina* [*Archives of Medicine*]; and in a book in which he applied some of his ideas to the world of culture, society and education, *Sketches of Social Pathology and Ideas on General Pedagogy* (1901). It is possible that had it not been for his untimely death at fifty, he had devoted more attention to the problems of explaining the presence of consciousness in the biological world.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

The theory of Lacerda is, by its originality, worthy of study. The representation of a future moment in time when humanity will feel nothing and nothing will have subjective awareness, but will live as if it were an automaton is undoubtedly worthy of reflection. From the point of view of this nineteenth-century author, the process of brain automation is inevitable and at the same time desirable. If evolution proceeds in this way, it follows that living without consciousness is a perfection that people should desire. It is the liberation of a burden that diminishes human beings.

Undoubtedly, it was in his 1895 book that the question of the future automatism of the human brain began to be analyzed in depth. More interestingly, Lacerda asked one of the great masters of Portuguese medicine at the time, Dr. José Tomás de Sousa Martins (1843-1897), to write the preface to this book.[1] One knows what is usually expected of such a document: a complimentary speech about the prefaced work or its author. None of this happened. Sousa Martins goes beyond what is recommended by the protocol of the writing of prefaces and strongly criticizes the idea that human consciousness will disappear in the future. In doing so, he throws down the theory in the very book in which the theory is presented. What was at stake in this clash of opinions in the unlikely place of a preface, which in principle should support the philosophical ideas of the author of the book? By carefully reading these surprising texts from the history of the nineteenth-century representations of mental life and theories of consciousness in the natural order, it is possible to find a war of ideas which to a large extent has not yet ended. This war of ideas is an important chapter in the intellectual history of the problem of consciousness in the physical world.

II

What, then, are the arguments that support the thesis that the human brain will be automatic in the future? José de Lacerda begins by drawing an analogy between what happened with the evolution of the brain and what happened with the evolution of the brain stem (*medulla oblongata*). This was the most complex nervous organ before the appearance of the brain, as seen in protozoa such as *ascidia* and *amphyoxus*. For Lacerda,

[1] On the scientific role of this doctor's work, see Repolho, 2008; on the vast influence he exerted in Portuguese society, to the point of deserving a religious cult after his death, see Pais, 1994.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

this organ was organized, developed, specialized and eventually automated, that is, functioning perfectly in the most complete absence of consciousness. From this sequential verification, Lacerda draws this conclusion: what has happened to the brain stem will also happen to the brain.

The equation of the problem is as follows. He asks himself: "Why did the nervous system – which has always been, from the first protozoan to the last acranial [*sc*. brainless animal], more or less conscious, as seen in the present representatives of the extinct ancient animals – become functionally automatic at all when it defined itself morphologically in brain stem and spinal nerves?" (1895, p. 22). There are two main ideas in this question. The first is that the simpler beings have a great sensitivity to the environment, feel and respond to the stimuli; they are therefore aware. The second idea is that, at a later point in the evolutionary process, the functions of the brain stem (medulla) ceased to be accompanied by consciousness, that is, they became unconscious. The core of his thought is thus the following: for a long time, simple living beings experienced sensitivity and awareness, responding to the demands of the environment; the biological development over the millennia of the phylogenetic evolution made the brain stem cease to be conscious and became automatic. As a result, the consciousness that has been associated with brain activity in the process has come to be regarded as an ephemeral side effect of the activity of this new organ, which, sooner or later, will also disappear, just as it did with the brain stem. This sequence of ideas relies heavily on a reconstruction of evolutionary periods very backward in time, and also depends on the conviction that in the nineteenth century there are still representatives of animals already extinct. An important part of the argument depends on the phenomenology of human activities, such as learning new tasks; however, in order to bypass the difficulty deriving from someone at present time proposing theories about a very distant moment in the past and on the evolutionary scale, Lacerda's thought requires that the past biological beings have "representatives" at the moment that person reflects on the conscious minds of the ancestors of these beings. Of course, the problem of other minds (how do we know that some other being is feeling something?) still exists, but the existence of representatives in the present time of past beings attenuates the difficulty of demonstrating the continuity of consciousness in all the moments that connect the past to the present time. These argumentative constraints certainly weaken Lacerda's conjecture. By the very nature of the realities in question, it is not possible to discover material evidence of subjective experiences or about the level of sentience of beings that lived millions of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

years ago. Sentience leaves no archaeological remains by definition, and it is not easy to demonstrate beyond all reasonable doubt that the remains of organic structures indicate that these living beings had conscious experiences. In addition, the idea that ancient beings still have living representatives is especially fragile because it fails to answer many questions: proof of biological continuity between ancient and contemporary living beings; the still more difficult proof of the continuity of the mental experience between the past living beings and the contemporary living beings who are their alleged representatives (consciousness could have, for example, a discontinuous presence over millions of years); the not yet demonstrated notion that the context is irrelevant to these contemporary representatives of ancient beings, that, for example, the fact of living in a world where other beings have high levels of awareness is irrelevant to their existence, not having created to them an evolutionary pressure so that they, too, could experience consciousness; etc. One might ask, giving an opportunity to Lacerda's argument, why has evolution caused Nature to alter the properties of the brain stem, making it an organ without consciousness, and shifting the location of mental life to a more recent structure, the brain? After all, if the brain stem has reached a state of automatic and unconscious perfection, it is interesting to know what caused a "new, *conscious*, and therefore, *perfectible* organ" (1895, p. 23). Lacerda's answer is that, as living beings become more complex and occupy new environments, they need a body that allows the exploration of these new contexts. The alleged state of unconscious perfection was, after all, imperfect; the evolutionary pressure caused that the structure of the brain stem had to be complemented by another biological structure that allowed the survival of the individuals. If a perfect and unconscious structure already existed, the new structure could be biologically more complex but be devoid of sentience.

The brain is like the deep roots of trees, whose tip moistens to more easily make its way through dozens of feet of earth and stone. Dr. Lacerda not only considers the brain as the organ that is exploring the world, but within the brain itself there are zones that are more conscious and others that are less conscious until they become fully automatic. Lacerda, with his incipient knowledge of the structure of the brain, which was possible for a doctor of the late nineteenth century, distinguishes the gray matter of the brain, which is more dynamic and conscious, of the already automated white matter. It was not possible at the time to have enough information to account for the degree of automation of these parts of the brain.

Despite this gap, one can see that Lacerda tries to overcome this difficulty with other lines of analysis. In addition to this general

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

consideration of the two parts of the brain, there is still a need to see the subject at a lower level. The plastic and functional evolution of the nervous tissue gave rise to the consciousness that the organism has of itself and of the environment; however, the improvement of memory implied a decrease in consciousness. The structures of memory fix information and make the need for consciousness less urgent.

The two organs, the brain stem and the brain, are separated in the argument of Lacerda. Thus, "the brain stem is, from the archaeolithic age, an anatomically and physiologically perfect organ2 (1895, p. 24). In opposition to this, he states that "the brain, on the other hand, is structurally and functionally confused ... more in the cortical gray matter – which is the most evolutionary, the most modern, and the most conscious – than in the white internal mass ... relatively automatic" (1895, p. 24). And he concludes the comparison between the two organs by saying that "the brain is, therefore, in the animal series, an imperfect, hesitant organ in essays" (1895, p. 25).

This is a very interesting interpretation of the evolutionary process. Lacerda argues that consciousness plays a useful role but is doomed to disappear by the normal order of things. The structure of the brain is itself an example of the processes that took place in the long time of evolution. What has happened in the transition from the brain stem to the brain is happening again within the brain itself: some parts are already fully automated, that is, their activity is not accompanied by consciousness. To the approach of the phenomenology of the learning of new tasks, a theoretical proposal on the architecture of the human brain is added. These two lines of reflection are inseparable.

This interpretation of the evolutionary past of mental life is epistemologically difficult to prove. However, the stroke of genius of Lacerda was to associate the reflection on the past of evolution on planet Earth to his reflections on how people learn new tasks, and how people perform tasks after many years of practice. Just as ancient living beings have representatives in contemporary living beings, and just as the interior of the human brain shows a process that has already occurred in other evolutionary periods, so too the behavior of the more developed contemporary beings is full of lessons. At the beginning of learning new tasks, or of the knowledge of new environments, the consciousness of the individuals seems to be especially intense. Tasks are performed with little perfection, as if consciousness itself were an obstacle to their accomplishment. After a long time of practice, people perform these same tasks with a much lower level of consciousness and sometimes even in a totally unconscious fashion. Lacerda performs very detailed analyzes of

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

this process with the pianists, the academics of the natural sciences and with the mathematicians. Being the brother of the composer, pianist and orchestra leader Francisco de Lacerda, he is likely to have made many of these observations in his own home environment, seeing how a musician's consciousness changes between the moment he begins to study a score and the moment that he performs it almost unconsciously. He thus summarizes the observations on the process of human learning as manifested in the performance of a musical instrument: "all our acts are less imperfect and costly the closer they approach unconsciousness, the more they become attuned to automatism" (1895, p. 25).

Alongside the study of the process of learning new tasks, José de Lacerda adds further analysis. Sleep without dreams, for example, seems to be an anticipation of the perfect life that will be achieved when all human life is unconscious. Just as the human brain already reveals that some parts are fully automated and others still suffer from consciousness, and just as the learning process reveals that habit can lead to automation of tasks, sleep is an example of how, at the level of macroscopic behavior, the human being can already live absolutely automated, without the defect of the hesitant consciousness. Lacerda says that

> In complete sleep, serene, healthy, without dreams, in which the phenomenon of consciousness is completely abolished, the nervous system does not cease to perform the reflexes associated with circulation, breathing, etc. These nervous, rhythmic, just, physiologically perfect acts performed without the slightest intervention of consciousness represent the most complete type of automatic reflexes. On the contrary, conscious reflexes – such as the complete series of reflexes unrolled in an individual who, for the first time, and with the utmost attention, performs a scale on a piano – demand maximum vigilance (1901, p. 26).

Lacerda's various lines of thought are summarized unambiguously by himself. On the one hand, he establishes the basis of the conjecture on the difference between the stem brain and the brain: "The stem brain was conscious, but it is and will be automatic; the brain has been and is conscious, but it will come to automatism" (1895, p. 27). On the other hand, trying to understand the provisional role of consciousness, it states that it is "a childish and decreasing way of neural sensibilities, necessary for the attainment of education; is an inferior and transient state of nervous receptivity, indispensable for the attainment of automatism" (1895, p. 27). The phenomenology of learning new tasks, either by self-observation or by observing other people, is supplemented by reflections on the structure of the brain.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Observations about altering the intensity of consciousness throughout the learning of complex tasks are insightful and intellectually promising. One sign of this is the fact that some of these conjectures and analyzes were taken up later by other researchers. Half a century after Lacerda, for example, Erwin Schrödinger will propose a similar reading of the connection of consciousness to learning in his Tarner Lectures, given at Trinity College, Cambridge University, in 1956. To Schrödinger, there are three groups of processes in human biology: a) monotonous processes that do not require conscious decisions and do not depend on the environment, such as heart beats and peristaltic movements; b) processes that occasionally require conscious decisions, such as breathing in risky atmospheres; and c) everyday processes, linked to habit and innovation. To Schrödinger, only the processes that are still being tried are conscious; long after these first trials, they become an unconscious heritage of the species that is hereditarily established (Schrödinger, 1985, pp. 14-18; 2002, pp. 95-99). Summing up with literary elegance his thought, Schrödinger states that "consciousness is the tutor who supervises the education of the living substance, but leaves his pupil alone to deal with all those tasks for which he is already sufficiently trained" (2002, p. 97). In a clearly Jamesian spirit, Schrödinger argues for the effective role of consciousness in the action of individuals, being a factor that favors their biological survival in a fiercely competitive world. Lacerda inserts himself precisely in this line of thought: consciousness plays a useful role as long as it exists, but this utility disappears when the processes it accompanies become absolutely automated. In short, consciousness seems to be a temporary moment, a detour in the process of biological development. This detour is structured through the representation of internal and external processes, and it is this representation that has the capacity to causally influence the behavior of the individual and, ultimately, his survival. William James states that consciousness acts as if it were an organ added to the biological organs of the body: "it [*sc*. consciousness] seems an organ, superadded to the other organs which maintain the animal in the struggle for existence; and the presumption of course is that it helps him in the some way in the struggle, just as they do" (1950, p. 138). Lacerda, who does not seem to have read James, develops a similar idea. He asks, trying to objectify the essence of consciousness: "What then is consciousness, as a psychic phenomenon?" He then replies: "It is the property that the *pallium*, or some part of the *pallium*, has to represent and influence the reflexes or stretches of reflexes, new or ill-known" (1901, pp. 28-29). This concept of consciousness emphasizes content; says nothing about the subjective aspect of consciousness; says also nothing about the precise location of the neuronal

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

system that can produce this representation, mentioning only the hypothetical role of the thin film covering the vertebrate brain (the *pallium*).

Observations about the evolutionary process in general are, as has been said, very fragile due to the epistemological problem of proof. However, the general features of Lacerda's argument seem plausible: simple organisms do respond to the environment, thus manifesting their consciousness; parts of the nervous system are losing plasticity and others are gaining new plasticity; habit seems to reinforce the somatic structures; the conscious reflex act is characterized by difficulty, slowness, hesitation, representation by images, mental vision prior to externalization by action, ability to causally influence the externalization that manifests itself in the behavior of the individual. These are intuitions that would be explored decades later by other researchers, such as Donald Hebb, with his reinforcement rule, and like Paul MacLean, with his triune brain theory.

The thinking of Dr. Lacerda has, therefore, a very strong aspect: the observation about human learning. It also has two more fragile aspects: the reconstitution of the evolutionary past and the foreseeing of the future of evolution. Taken together, these three parts can be considered an interesting conjecture about the role of consciousness in the order of nature. If this theory seems plausible and intellectually stimulating, to which one could add the fact that other authors proposed it during the twentieth century, why did the distinguished pathologist Sousa Martins refuse to accept it?

III

In the preface to the book *The Neurasthenics*, with a strong prose that still causes surprise, Sousa Martins is not a diplomat and immediately says what he thinks about the theory of the future automatism of the brain: it is a heresy that does not seem acceptable to him (1895, pp. XIX, XXII). His argument is concerned with the two main aspects of the question: the role that the brain plays in the body organs and the meaning of conscious mental life.

In regard to the former, the idea that the brain is an imperfect organ because of its alleged hesitations and essays is strongly criticized. For Sousa Martins, there is no possibility of absolute assertion that the brain is imperfect because perfection seems to be a relational and contextual property; it is not an intrinsic property to the organ, nor is it absolute. Moreover, Sousa Martins defends the idea that the brain is not an exception in the set of organs of the human body (1896, p. 224). All the

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

organs of the body are equally perfect and imperfect, depending on the context. The possibility that the brain may be better in the future than it is in the present cannot be considered an imperfection. Obviously, when one achieves superior development, by looking back to what existed in the past, the past moment can be considered to be manifestly imperfect. However, this alleged imperfection had sufficient resources to enable final perfection, and this could not have existed without the initial imperfection. Therefore, a process of improvement is in itself a sign of perfection. Sousa Martins himself generalizes the localized question of the brain to all other organs of the body: "From the fact of being perfectible, the imperfection is not deduced. It is, the brain, perfectible ... But this quality is common to all other organs" (1895, p. XXIII). Looking very simple, the thinking of Sousa Martins points to very complicated metaphysical questions about the relation between different states of the same temporal process; moreover, it forces us to think about the meaning of the perfection that is attributed to the different organs. Some somatic organs seem more developed; others, more primitive. Some seem atavistic and devoid of function, but may have had a function in the past; others seem essential to the functional economy of the body. Some seem more fragile than others, though this assessment depends on possible contexts. Some organs appear to be associated with mental structures, such as the nervous system, while others are devoid of this connection. Many other parallels could be established.

Questions about brain perfectibility can also be asked about other organs. The liver, for example, reveals many improvements throughout zoological history. In the phylogenetic past, it separated from the spleen and pancreas; in the future, there may be a "biliogenic" and a glycogenic liver, whether in man or in other higher species. For Sousa Martins, this is a concrete case of functional specialization that derives from the advantages of new somatic configurations. It is, he says, a "simple case of the law of division of labor. The liver has already profited greatly; much will profit in the future and more than probable specializations" (1895, p. XXIII). Even if one does not refer to the great temporal periods of phylogenesis, in the lifetime of individuals there is something similar. Structured activity contributes to improved organ performance. Sousa Martins gives examples of the improvement of the leg muscles due to the practice of mountaineering and the alteration of lung activity in high altitude contexts (1895, p. XXIII). This generalization to all other organs of the possibility of improving the brain is inspiring of other kind of generalizations. The case of humans can also be generalized to other

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

biological species. For Sousa Martins, Dr. Lacerda was therefore not seeing the problem in all its extension.

The sense in which Dr. Lacerda takes consciousness is also criticized. For Sousa Martins, Lacerda only considered the cerebral consciousness that allows the existence of personality and morality. This conscousness is, for Sousa Martins, very imperfect when dealing with healthy organs, and very misleading when dealing with diseased organs. For this physician, human consciousness is finite, very partial, and prone to deception and illusion. In addition to this brain-associated consciousness, one would have to consider other hypothetical consciousnesses. Sousa Martins proposes an extension of the concept of conscousness. The digestive system does not appear to be, in his view, very different from the brain. This last organ, for Lacerda, is characterized by hesitation, trials or attempts, that is, by the ability to adapt to situations. Sousa Martins is not impressed with the idea that the mutability of the brain is greater in the periphery, in relation to the senses of external perception, than in the more rigid parts of the brain, which regulate, for instance, respiratory and cardiac functions. The digestive system would also reveal these hesitations and trials if it were confronted with quantitative and qualitative changes in diet. It could even lose all this ability to adapt, and become a fully automatic system. Sousa Martins even speaks about a "gastric consciousness" (1895, p. XXIV). The stomach responds to food, adapts to variations in food, perceives and feels what it ingests. What is this other than having consciousness, or at least having a form of consciousness that goes unnoticed to personal consciousness? As a test for this conjecture, the cases of dissonance between the personal and the gastric consciousnesses may be mentioned: there are substances that, if they were consciously ingested, would provoke aversion and repugnance; taken unconsciously do not provoke such reactions. We have here a very interesting theoretical proposal for the extension of the concept of consciousness. It was seen that Dr. Lacerda anticipated in many decades later investigations on the automatization of the tasks by the effect of the habit; but it must also be acknowledged that Dr. Sousa Martins' critique anticipated proposals of modularity of mind by associating each somatic system with a certain degree of sentience. The case of the digestive system can be generalized to many others. Take the case of the marrow. A decapitated frog cannot be aware of its individuality because it has no head, but has a consciousness linked to its other organs, like liver and marrow. The frog is totally "*unconscious as a frog*, but conscious as a *living* marrow, sensitive marrow, judgmental and willing" (1895, p. XXV). In view of the brainless situation in which it is, the frog searches for the free limbs at its disposal to

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

make some movement. Of course, these free limbs have no access to the ideal end of the frog's movement, but neither did they when the frog still had a head. Sousa Martins, trying to explain the limitations of the consciousness of each of the non-cerebral organs, resorted to the image of a human army. The soldiers do not know what the headquarters know. In a sense, soldiers are automata who carry out orders without being aware of what motivated these orders or the scope and ultimate meaning of them.

The center of this debate is the notion of consciousness. Sousa Martins affirms that Lacerda did not consider the consciousness *of* the brain, that is, the consciousness that the brain has of itself as an organ, that is, a regional consciousness. From his point of view, Lacerda addressed only the consciousness *in* the brain, that is, the individual's awareness of his personality, or the organism's individuality. This latter form of consciousness is connected to the senses; it is dynamic because perception compels it. Consequently, one can only think of a future without consciousness in the brain in a scenario of disappearance of the senses. He says that if the senses were "pathologically abolished" (1895, p. XXVI), human beings will cease to be aware. This is not as fantastic a scenario as it may seem. Sousa Martins even considers that there may already have been individuals without senses due to diseases, and that, therefore, there is a serious possibility of this happening in the future to the whole human species. He mentions for sure that it depends on an "inconceivable cosmic change" (1895, p. XXIV), meaning that the possibility is manifestly remote, but it is not negligible. If, for the sake of argument, this happened, one would lose consciousness of one's self, for without senses one would have no way to compare oneself with that which is not self. He would speak of himself in the third person, lose his psychic consciousness, and be without personality. He would have no extrinsic impressions, though he remained alive. The abolition of the nerves of the common sensibility, of the senses and of the will, connected with the behavior, would imply the complete atrophy of the nervous system, perhaps even its disappearance. The animal would become the vegetable: "What, in this case, would become an animal species, without senses? It would be a living species, *without nerves*. Just a *plant* species" (1895, p. XXVI). It seems to be an even more serious human condition than the few cases in which humans lost many senses of external perception, such as the American ladies Laura Bridgman (1829-1889) and Helen Keller (1880-1968). The scenario described approximates the state in which humans could be considered unissensorial, that is, beings with only a single sense, or even without any animal sense of external perception, similar to plants. In this hypothetical scenario, the theory that the brain can become automatic would make no

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

sense. If the brain were to be non-existent due to the disappearance of the senses and the atrophy of the nervous system, it could not be said that what does not exist is at all automatic (1895, p. XXVIII). Accepting another scenario less extreme than this one, there are also reasons why it is not easy to understand how the future automatism of the brain could be possible. If, for the sake of argument, it was accepted the hypothesis that personality consciousness could disappear, as Sousa Martins proposes, what would happen next? In an analogy with the decapitated frog, its organs seem to maintain the consciousnesses that are unique to them. For example, there would still be a consciousness of the lymph node system. The parallel that Sousa Martins draws is of a political nature. In the case of losing consciousness as a federation, as a whole, there would still be a district consciousness and the "organism would be a federation of tiny ganglionic consciousnesses" (1895, p. XXVII). This line of argument based on possible scenarios does not correspond, however, to the idea that Sousa Martins has of the future of the evolution of the conscious mind. In a collision course with Lacerda, Sousa Martins believes that the future will see an amplification of the consciousness of human beings. His argument is *top down*, that is, he moves from top to bottom. The brain is extremely sensitive to variations in the environment. The great creators of culture force the brain to change and rise on the scale of "mental progression" (1895, p. XXX). The examples he gives of great culture are remarkable: Christ, Gutenberg, Columbus, Luther, Diderot, Watt, Lavoisier, Pasteur, etc. The intellectual proposal is very powerful: culture has the ability to influence the development of the brain. In the context of the preface to the book *The Neurasthenics*, it is not explained in detail how the *top* of culture actually changes something concrete such as the *down* of the physiology of the brain. The insights of Sousa Martins are, however, of great interest and approach to the investigations of his contemporary on the other side of the Atlantic, the American psychologist James Mark Baldwin (1861-1934), and can be considered as anticipating also very later theories that link development of the human brain and, above all, its size, to the effects of the creation of material culture (language, art, instruments, etc.). There is now a conviction that social complexity has been a decisive factor for increased memory, for brain growth and even for the theory of the mind module (in a complex social environment, there is an advantage for systems that anticipate the intentions of others). Sousa Martins' conviction that the brain is an "apparatus in growing nobility" (1895, p. XXIX) compels him to defend the idea that sensitivity to the environment is a positive factor. Unlike Lacerda, who sees this sensitivity as hesitation and weakness, Sousa Martins is in fact describing the virtues of the plasticity

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

of the brain. In his time, William James spoke of the brain as having a "hair-trigger organization", that is, a great sensitivity to the variations of the environment in such a way that responds to them quickly (1950, I, 140). In addition to the rapid response to the environment, the brain places cultural products in the environment, which in turn changes the context in which individuals move. Sousa Martins' argument describes a causal interdependence: culture and social complexity promote the increase of the size of the brains, and, consequently, the increase of the brain will imply a greater increase of the consciousness of the human beings. In conclusion, the future will have human beings with a level of intensity of consciousness far greater than what exists today. Sousa Martins ostensibly states that this is a process of amplification of consciousness that will reduce "the distance – in any case infinite! – which separates it from supreme perfection, from the *Absolute*, by each mythology incorporated into its respective Jupiter" (1895, p. XXX). Despite the caveat of the infinite distance between human and divine consciousness, there is no doubt that Sousa Martins is optimistically describing the unlimited perfectibility of human consciousness.

Lacerda, for his part, says that the human beings of the future will be automatic, that is, they will live without feeling anything, without consciousness. Generalizing the case of individuals to the totality of societies, the automation of societies would imply a life similar to that of insects and that of even simpler organisms. Lacerda's idea of social perfection is that of colonies of beings that do not seem to have individual consciousness. Against the most elementary of the evidences available, Lacerda criticizes the greatest of human achievements. In his own words, "Human societies, as organisms, are rudimentary ... much lower, for example, than earthworms" (1901, p. 57). Justifying this criticism, Lacerda denounces the lack of symbiotic mechanisms in the human world and the lack of a complete integration: "The social symbioses ... have not yet passed from childhood, have not yet left the diapers; they imitate, for the time being, in social perfection and in clear results, the simplest colonies of the most modest hydra" (1901, p. 57). As one would expect, this parallel between human societies and very simple animal societies justifies a wide range of criticisms that Lacerda makes to Belle Époque society. The symptoms of the mal-de-vivre are many: boredom, indifference, war, nihilism, anarchy, hypossociability, hipobulia, etc. (1901, pp. 75, 82, 84). Lacerda also sees a decrease in the sensory activity of humans in modern cities. This atrophy of sensory acuity (hyposensoriation) will tend to increase and to generalize, and is interpreted as a sign of progressive loss of consciousness (1901, p. 157).

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

IV

These two authors could not, as can be seen, affirm more radically antagonistic theses. What, then, can one conclude from this conspicuous difference of theses on mental life and its role in the natural order? The future automatism of Dr. Lacerda's brain is not the same as the loss of personal consciousness due to the atrophy of the nervous system, a hypothesis granted by Sousa Martins. The level of confrontation and divergence of opinion is surprising in a place that is not the most appropriate to express differences of opinion. Surprisingly, above all, both authors have not noticed how their perspectives share common problems.

i. Lacerda proposes that the problem of consciousness can be equated in two ways: that of biological species and that of human individuals. On the one hand, it states that consciousness exists at the beginning of the biological process (in biological species with a simple structure) and tends to disappear as it rises on the biological scale. On the other hand, consciousness is especially intense at the beginning of the actions of the individuals and disappears as the novelty of the situations is being dominated. The common element to species and individuals is the existence of consciousness in the initial moments and the tendency to disappear in later moments. The problem is that both authors do not explain where consciousness comes from. Is it caused by brain activity? This does not seem to be the case, because even the brainless beings, such as amoeba, algae and fungi, are sensitive to the environment. On his side, Sousa Martins points to an explanation: each level of complexity of matter has its own consciousness, being possible to speak of an atomic, molecular and cellular consciousness, which is, of course, a way of affirming pampsychism, that is, the idea that consciousness is common to all matter. Nor is it explained where comes from consciousness that exists in all levels of organization of matter, or the differences between sentient matter and inanimate matter (at least, apparently).

If we ignore Lacerda's initial problem and the problem of Sousa Martins' pampsiquism, one can see that there may be no radical incompatibility between the two. Lacerda states that the mental and cerebral systems become automatic, but also recognizes the need for awareness to accompany the onset of tasks and the adjustment to new environments. Sousa Martins would affirm the same in a different way: the new environments create the necessary situations to develop the amplitude of human consciousness.

ii. The divergence between them is intellectually stimulating. The theses that are at stake can be used with benefit by other researchers.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Sousa Martins denounces the referents of the word 'consciousness' used by Lacerda. Two aspects of this problem are especially relevant: the identification of conscious activity in other non-human species and the identification of this activity in other humans. He says, "what processes would we resort to ... to make sure of the reality of consciousness in the other members of our species, and what processes lead us today to the recognition of consciousness in species mute for us?" (1895, p. XXXI). As is easily perceived, is at stake the proposal for a criterion for evaluating the presence of conscious activity. If one perceives the need for a criterion, it is clear that some form of criterion would have to be proposed. Sousa Martins seems to point to the tiniest degree of action: the choice or decision. The boundary that separates the inanimate from the animate seems to be here: "All activity involving selection translates into a consciousness" (1895, p. XXXII).

iii. There is a common element to the perspectives of the two doctors, an element that none of them realized. It is this: none of them noticed that it is surprising that any biological structure, whether or not it has a nervous system, is accompanied by sensations or degrees of consciousness. Biological beings, however simple or complex they may be, could live unconsciously. Lacerda and Sousa Martins have both sought to assign a function to the phantom of conscious sensations in such a way that the possibility of consciousness being an epiphenomenon without causal influence in the individual's life could be definitively ruled out. If consciousness has a function, then it serves something. In turn, if it serves anything, then one cannot think that biological organisms could eventually become unconscious automata. If consciousness has a function, then it is useful for something. In turn, if it is useful for something, then one cannot think that biological organisms may in the future become unconscious automata.

To conclude, the following must be said. The biographical circumstances of the authors who contribute to a scientific problem with arguments, concepts, conjectures and modes of interpretation are usually devalued. It is important the substance of scientific questions and not the historical framework in which they were produced. However, the present controversy deserves to be treated differently. In the volume *In Memoriam* dedicated to Sousa Martins, Lacerda writes a funeral eulogy to the author of the preface that undermined the credibility of his own theory. It's such a fair text that it even seems unfair. Sousa Martins is analyzed coldly by Lacerda who makes considerations about the physiognomy, intellectual formation and abilities of the author of the *Nosography of Antero*. Sousa Martins, for Lacerda, had a "strange, irregular physiognomy", was

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

"imposingly ugly" and even showed traces of "Ethiopic atavisms" (1904, pp. 301). In considering the great clinician' writing, he coldly states that "the little he wrote [*sc*. Sousa Martins] is thwart and inferior" (1904, p. 306). It is probable that this coldness disguised by the appearance of justice is a sign of a divergence of opinions that has never disappeared.

Be that as it may, this intellectual debate between two friends compels their readers to think about what it means to be aware. One can only be grateful to these clinicians who had manifest interest in philosophical matters. The problems that enchanted them remain unresolved more than a century later. An unresolved philosophical problem makes each chapter of its intellectual history precious. As if it were a map of the path already covered, intellectual history is the only possibility to see clearly where reflection was gone wrong and why it is taking so long to find a solution.

References

Curado, Manuel (2012). "A descoberta do inconsciente no século XIX português [The discovery of the unconscious in the Portuguese 19th century]", *Diacrítica/ Filosofia*, 26: 2, pp. 157-182.

James, William (1950). *The Principles of Psychology*, 2 vols. New York: Dover Publications [1st ed., 1890].

Lacerda, José Caetano de Sousa e (1895). *Os Neurasténicos: Esboço de um Estudo Médico e Filosófico* [The Neurasthenics: Outline of a Medical and Philosophical Study]. Pref. Sousa Martins. Lisboa: M. Gomes.

Lacerda, José de (1896). "A consciência e o livre-arbítrio [Consciousness and free will]", *Medicina Contemporânea*, p. 102.

—. (1897). "Hipnologia [Hypnology]", *Arquivo de Medicina*, I, pp. 60-63, 124-128, 176-179, 399-403, 573-580.

—. (1901). *Esboços de Patologia Social e Ideias sobre Pedagogia Geral* [Sketches of Social Pathology and Ideas on General Pedagogy]. Lisboa: Livraria de José A. Rodrigues.

—. (1904). "Um Homem [A Man]". In *Sousa Martins (In Memoriam)*. Lisboa: s.n., pp. 301-310.

Martins, José Tomás de Sousa (1895). "Prefácio [Preface]", in José Caetano de Sousa e Lacerda, *Os Neurasténicos: Esboço de um Estudo Médico e Filosófico*. Lisboa: M. Gomes, 1895, pp. V-XLIV.

—. (1896). "Nosografia de Antero [Antero's Nosography]", in *Antero de Quental In Memoriam*. Porto: Mathieu Lugan Editor, pp. 219-314.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Pais, José Machado (1994). *Sousa Martins e suas Memórias Sociais: Sociologia de uma Crença Popular* [Sousa Martins and his Social Memories: Sociology of a Popular Belief]. Lisboa: Gradiva.

Repolho, Sara (2008). *Sousa Martins: Ciência e Espiritualismo* [Sousa Martins: Science and Spiritualism]. Coimbra. Imprensa da Universidade de Coimbra/Coimbra University Press.

Schrödinger, Erwin (1985). *Mente y materia. Conferencias Tarner leídas en el Trinity College, Cambridge, en octubre de 1956*. Transl. Jorge Wagensberg. Barcelona: Tusquets [1st ed., 1985].

—. (2002). *What Is Life? The Physical Aspect for the Living Cell* with *Mind and Matter & Autobiographical Sketches*. Cambridge: Cambridge University Press [1st ed. 1944 and 1958].

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

CONTRIBUTORS

José Antunes was born in 1963 in Benavente, Ribatejo. Lisbon was the stage of his basic training at the Faculty of Letters in Portuguese Studies. In 1989 he moved to Porto and was responsible for promoting the Nova Acrópole school of philosophy. At the Faculty of Letters of Porto he studied History with a focus in archeology. He continues his training at the Nova Acrópole school of philosophy, teaching and lecturing various conferences, mainly on historical and philosophical topics. He has written and published several texts, all in Portuguese: "A Filosofia e o Despertar da Consciência Histórica", "A Ordem de Santiago em Portugal", "A Verdadeira Felicidade", "Um Santuário nas Terras de Ofiúsa – Castro do Baldoeiro, Um castelo que nunca foi – Castelo do Mau Vizinho", "Traição e morte em Rodrigues Lobo", "Os Ciclos na História e os Leitmotifs Humanos", "O Teatro Grego", "Francisco de Holanda", "Dante e a Divina Comédia" and "Cachão da Rapa: um Enigma da Arte Rupestre".

André Zamith Cardoso is a researcher and technology analytical chemist at Innospec Inc. He completed a BSc in Biomedical Engineering and MSc in Bionanotechnology. He is completing a PhD in Chemistry at the University of Liverpool, UK. His interest in neurosciences and his previous work with human-machine interfaces at the Telecommunications Institute (Lisbon, Portugal) emerge as part of his optional studies, complementary to his focus in understanding matter interactions at a supramolecular scale (Soft Matter, (12) 3612-3621, 2016).

William Child is a professor of Philosophy at the University of Oxford. He is the author of *Causality, Interpretation, and the Mind* (Oxford University Press, 1994) and *Wittgenstein* (Routledge, 2011), and has published widely on issues of philosophy of mind and on Wittgenstein.

Judite Maria Zamith Cruz has a PhD in Psychology and is a professor at the University of Minho, Portugal. She has training in educational psychology and a master's degree in Philosophy, having specialized in cognitive sciences (J. Zamith-Cruz, Apeiron, (4) 31-38, 2014; J. Zamith-Cruz, Apeiron, (6) 53-77, 2015). Her research is focused on neurodevelopment, disorders in

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

childhood and adolescence (Emotional and Behavioural Difficulties.16 (4) 419-436, 2011), sexuality and emotions.

Manuel Curado is a professor at the University of Minho, National Defense Auditor, Doctor *cum laude* from the University of Salamanca, MA from the Nova University of Lisbon, graduate from the Universidade Católica Portuguesa (Lisbon) and holder of the Senior Management Course for Public Administration (CADAP). He was visiting professor in the universities of Moscow, Russia (MGIMO and MGLU) and professor Erasmus in the University of Padova (Italy); he collaborated with the universities of Porto, Coimbra, Universidade Católica Portuguesa and Vigo. Moreover, he is the author of various books, all in Portuguese: *As Viríadas do Doutor Samuda* (Coimbra, Imprensa da Universidade de Coimbra, 2014), *Um Génio Português: Edmundo Curvelo* (Coimbra, Imprensa da Universidade de Coimbra, 2013), *Porquê Deus se Temos a Ciência?* (Porto, Fronteira do Caos, 2009), *Direito Biomédico: A Legislação Portuguesa* (Lisboa, Quid Juris, 2008), *Luz Misteriosa: A Consciência no Mundo Físico* (Famalicão, Quasi, 2007) and *O Mito da Tradução Automática* (Braga, Universidade do Minho/Cehum, 2000). Finally, he is also editor of several books: *Obras Completas de Edmundo Curvelo* (Lisboa, Fundação Calouste Gulbenkian, 2013), *Deus na Universidade: O que Pensam os Universitários Portugueses sobre Deus?* Prefácio de D. Jorge Ortiga (Porto, Fronteira do Caos, 2011), *Cartas Italianas de Verney* (Lisboa, Sílabo, 2008), *Pessoas Transparentes: Questões Actuais de Bioética* (Coimbra, Almedina, 2008), and two titles in collaboration with Alfredo Dinis, SJ, *Mente, Self e Consciência* (Braga, Universidade Católica Portuguesa, 2007) and *Consciência e Cognição* (Braga, Universidade Católica Portuguesa, 2004).

Luca Forgione is an associate professor of Philosophy and Theory of Languages at the Department of Humanities of the University of Basilicata (Italy). His research focuses primarily on philosophy of language, philosophy of mind and self-knowledge, with particular reference to Kant. His book *Kant and the Problem of Self-knowledge* is forthcoming.

Bryan Frances is a visiting full professor at Lingnan University in Hong Kong. He works in metaphysics, epistemology, philosophy of religion, philosophy of mind, and philosophy of language. He is the author of three professional books and about three dozen articles. He is also finishing up four philosophy books for a general audience.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Klaus Gärtner is currently a Post-Doc at the Center for Philosophy of Science of the University of Lisbon. His recent work revolves mainly around topics in philosophy of mind and cognitive sciences, philosophy of science, epistemology and metaphysics. In his recent PhD-thesis *From Consciousness to Knowledge - The Explanatory Power of Revelation* he defended that our privileged epistemic position in relation with our conscious experiences can be justified by the controversial revelation thesis. Lately, he is working on issues related to Enactivism, especially the metaphysical and epistemological implications of this view for cognition and experience.

Steven S. Gouveia studied Philosophy (with focus on Philosophy of Mind) at the University of Minho. He is currently working on his Ph.D project (under the supervision of Dr. Georg Northoff and Prof. Manuel Curado) on the methodological problem between philosophy and neuroscience, analyzing the several approaches (isolationist, neurophenomenology, reductive and non-reductive neurophilosophy) that try to deal with the problem. He will apply the best methodology in two central concepts of philosophy and neuroscience, qualia and information. Moreover, he is the organizer of the International Conference on Philosophy of Mind, held at the University of Minho. Finally, he is the editor-in-chief of a student journal of philosophy called *Apeiron*, where articles and papers of current and past students are published. *Apeiron* also has a special section with the presence of an honoured guest. Some issues of the journal are: no. 6, dedicated to "Philosophy, Literature and Cinema", with the contribution of Noël Carrol; no. 7, on "Philosophy, Computation and AI", with Daniel Dennett; and no. 8, on "Philosophy, Ethics and Animal Rights", with Peter Singer. Upcoming publications will have the participation of Slavoj Žižek (no. 9) and Noam Chomsky (no. 10). The journal is available on Amazon. His current research interests include some of the big questions of philosophy of mind, the relationship between technology and society – broadly understood – and the definitions of art.

Matt Mahoney is retired now. He got his Ph.D. in Computer Science at Florida Tech in 2003 with a particular interest in AI. He was a chief scientist at Ocarina (acquired by Dell) from 2008 to 2015 conducting research in data compression algorithms and their relationship to machine learning. His work and papers can be found at http://mattmahoney.net/.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Thomas Metzinger is a professor of Philosophy at the Johannes Gutenberg-Universität Mainz and an Adjunct Fellow at the Frankfurt Institute for Advanced Studies. He is past president of the German Cognitive Science Society and of the Association for the Scientific Study of Consciousness (from 2009 to 2010). He has edited two collections on consciousness *("Conscious Experience"*, Paderborn: mentis & Thorverton, UK: Imprint Academic, 1995; *"Neural Correlates of Consciousness"*, Cambridge, MA: MIT Press, 2000) and one major scientific monograph developing a comprehensive, interdisciplinary theory about consciousness, the phenomenal self, and the first-person perspective (*"Being No One – The Self-Model Theory of Subjectivity"*, Cambridge, MA: MIT Press, 2003). An important recent open access publication is Open MIND at open-mind.net. In 2009, he published a popular book, which addresses a wider audience, discussing the ethical, cultural and social consequences of consciousness research (*"The Ego Tunnel – The Science of the Mind and the Myth of the Self"*, New York: Basic Books).

Sofia Miguens is a professor at the Department of Philosophy of the University of Porto. Her MPhil and PhD studies were supervised by Fernando Gil (École des Hautes Études en Sciences Sociales (EHESS)), Paris / Nova University of Lisbon (FCSH/UNL). She was a visiting scholar at New York University (Fall 2000), a visiting research fellow at Institut Jean Nicod-Paris (2007-2008) and a visiting scholar at the University of Sydney – Austrália (2013). She was President of the Portuguese Philosophical Association (Sociedade Portuguesa de Filosofia, SPF) (2004-2006) and head of the Philosophy Department at FLUP (2008-2010). She was a member of the board of the Institute of Philosophy of the University of Porto from 2007 to 2014, and currently sits in its scientific committee. She is the founder and principal investigator of MLAG (Mind, Language and Action Group), a research group of the Institute of Philosophy of Porto. She is the author of six books (*Uma Teoria Fisicalista do Conteúdo e da Consciência – D. Dennett e os Debates da Filosofia da Mente* [2002]; *Racionalidade* [2004]; *Filosofia da Linguagem – uma Introdução* [2007]; *Será que a Minha Mente Está Dentro da Minha Cabeça?* [2008] and *Compreender a Mente e o Conhecimento* [2009] *John McDowell – uma Análise a partir da Filosofia Moral* [2014]) and editor of several others, among them *Aparência e Realidade* (2010), *Aspectos do Juízo* (2011), *Acção e Ética* (2011), *Consciousness and Subjectivity* (2012) and *Pre-reflective Conscioueensss – Sartre and Contemporary Philosophy of Mind*. She has published numerous articles, review articles and book chapters in Portuguese, English and French on several topics in philosophy

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

of mind, language and action; epistemology and cognitive science; and history of 20th century philosophy.

Jaime Milheiro is one of the most renowned Portuguese psychiatrists and psychoanalysts. He was president of the College of Specialty of Psychiatry of the Order of Physicians for six years (1981-87). Furthermore, he was a visiting professor at the Faculty of Psychology of the University of Porto, between 1977 and 1982. He was also the president of the Portuguese Association of Mental Health. He is a full member of the Portuguese Society of Psychoanalysis and the International Psychoanalytic Association since 1981. He is also a lecturer at the Institutes of Psychoanalysis in Lisbon and Porto. Moreover, he was president of the Portuguese Society of Psychoanalysis (1990-92) and director of the Portuguese Journal of Psychoanalysis (1995-2003). He has written dozens of scientific articles and hundreds of opinion pieces in collective books, magazines and newspapers, often expressing personal ideas and proposals for the future.

Diana Neiva is a master's student in Contemporary Philosophy at the University of Porto. She has started writing her thesis on philosophy through film and metaphilosophy in 2016, under the supervision of Professor Sofia Miguens and Professor Thomas Wartenberg. She graduated in Philosophy at the University of Minho, where she organized various sessions of "Philosophical Cinema" and an International Conference on Philosophy of Mind. Her current research interests include metaphilosophy, film as philosophy, the cognitive approaches of film spectatorship, horror film, feminist theories of film, and philosophy of mind.

Eray Özkural has received his PhD in computer engineering from Bilkent University, Ankara. He has a deep interest in the philosophical foundations of artificial intelligence, and has proposed interpreting non-reductionism in philosophy of mind as epistemological irreducibility using concepts from algorithmic information theory. He has proposed a new approach to hypothesizing about subjective states of brain simulations within a strongly physicalist framework, suggesting that radical physical changes in a simulation are likely to result in a radically new form of qualia. He has also proposed a new foundation for information theory based on quantifying the exact physical resources required to produce data.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

Alfredo Pereira Jr. got his PhD from University of Campinas (UNICAMP; 1994) and was a post-doctoral fellow at the Department of Brain and Cognitive Sciences of the Massachusetts Institute of Technology (1996-1998). He published/organized four books and more than a hundred papers and chapters in the fields of philosophy of science and theories of consciousness. His original work on triple-aspect monism was first published in A. Pereira Jr. & D. Lehmann (Eds.) *The Unity of Mind, Brain and World: Current Perspectives on a Science of Consciousness* - Cambridge University Press. His most cited paper is on neuro-astroglial interactions: Pereira Jr, A. & Furlan, F. A. (2010) "Astrocytes and Human Cognition: Modeling Information Integration and Modulation of Neuronal Activity". *Progress in Neurobiology*, 92 , 405– 420.

Aleksandar Risteski is a Ph.D. student of Philosophy, at the University of Novi Sad, Serbia, where he received his bachelor (2014) and master (2015) degrees in Philosophy. His fields of interest are ancient philosophy, metaphysics, epistemology, pragmatism, philosophy of mind and philosophy of science. He has written book reviews and published articles on Plotinus, neuropsychoanalysis, pragmatism and phenomenology. He is also the main editor of the volume of the III Program, dedicated to Plotinus' Anthropology.

João de Fernandes Teixeira holds a B.A. in Philosophy (University of São Paulo, Brazil) and a PhD from Essex University, England. In 1998 spent a year as a visiting-scholar at the Center for Cognitive Studies of Tufts University, with the supervision of Daniel Dennett. He pioneered research in cognitive science and philosophy of mind in Brazil. Author of 14 books (some are now being translated into English) and many papers in national/international periodicals.

Keyvan Yahya is currently a researcher of neuroscience and applied mathematics at Chemnitz University of Technology, Germany. After receiving a BSc in Mathematics from University of Isfahan, he pursued his studies in cognitive neuroscience at University of Birmingham. He conducts his reach through a broad variety of topics ranging from mathematical modelling of perception and decision-making to consciousness. Besides, he has studied music perception from computational and cognitive perspectives which winded up in a number of studies in regards to understanding the cognitive underpinnings of Persian music. Since few years ago, he has begun concentrating over free energy and active

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use

inference as well as visual template learning which is considered to be an open problem since the late 1960's.

Benjamin D. Young is an assistant professor of Philosophy at the University of Nevada, Reno. He conducts research at the intersection of cognitive neuroscience and philosophy with a particular emphasis on olfaction. Most recently he has published articles on non-conceptual content, qualitative consciousness in the absence of awareness, and the perceptible object of smell. His current projects concern non-conscious phenomenology, tracking the senses, and the aesthetics of perfume.

EBSCOhost - printed on 2/14/2023 11:19 AM via . All use subject to https://www.ebsco.com/terms-of-use