Mind, Mechanism and Meaning

Faculty of Arts
Department of Philosophy

Mind, Mechanism and Meaning

Reclaiming Social Normativity within Cognitive Science and Philosophy of Mind

Thesis for the degree of Doctor in Philosophy at the University of Antwerp to be defended by

Farid Zahnoun

Supervisor: Prof. Erik Myin

Antwerp, 2018

Faculteit Letteren en Wijsbegeerte
Departement Wijsbegeerte

Cognitie, mechanisme en betekenis

Pleidooi voor het socio-normatieve binnen cognitieve wetenschap en filosofie van de geest

Proefschrift voorgelegd tot het behalen van de graad van doctor in de Wijsbegeerte aan de Universiteit Antwerpen te verdedigen door

Farid Zahnoun

Promotor: prof. dr. Erik Myin

Antwerpen 2018

Voor mijn moeder, Els.

For it is hard to believe that the subjective state of feeling fulfilled could occur without there being at least the belief that one is connected with some objective value.

—Arnold Burms

Contents

General Introduction

According to a longstanding iconography in Western culture, a philosopher is someone with his (!) head in the clouds, interested only in what lies beyond, therefore missing what is right in front of him. This enduring caricature is assumed to find its earliest incarnation in Thales of Miletus. To this day, philosophy undergraduates are still familiar with the story about the pre-Socratic thinker, one day accidentally tumbling into a well because of his preoccupation with things 'up there'. In Plato's *Theaetetus* – probably the earliest written version of the anecdote – we read:

> While he [Thales] was studying the stars and looking upwards, he fell into a pit, and a neat, witty Thracian servant girl jeered at him, they say, because he was so eager to know the things in the sky that he could not see what was there before him at his very feet. *The same jest applies to all who pass their lives in philosophy*. (Theaetetus 174 A, my addition and emphasis)[1]

The Thales anecdote itself appears to have been something of a running gag already in Plato's days. However, it is important to note that Plato doesn't merely repeat the anecdote. As the last sentence of the above citation makes clear, through the mouth of Socrates, he explicitly *endorses* the generalization of the caricature to "all who pass their lives in philosophy". Ironically, then, although the Thales anecdote itself predates Plato, the actual stereotype of the absent-minded philosopher may very well find its origin in the latter's philosophical writings.

In stark contrast to the lofty image of the philosopher stands that of the modern scientist. To be sure, both philosophers and scientists alike are to this day popularly conceived of as gray-haired, often bearded males, usually suffering from a severe lack of social skills. Yet, when it comes to their shared occupation – the search for knowledge and truth – the difference could hardly be greater.

[1] In *Plato in Twelve Volumes*, Vol. 12. Translated by Harold N. Fowler. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1921.

Instead of the speculative musings of the woolly-headed philosopher, the popular iconography of the scientist emphasizes the rigidity of the scientist's methods and his overall 'down to earth' attitude. His field is not that of intangible ideas, but of measurable facts. The white toga is traded in for the white lab coat, since facts are things that are discoverable, not so much by the power of the mind alone, but by carefully set-up experiments and well-calibrated measuring devices. The modern scientist doesn't tumble in wells, he studies them, for instance to learn more about the local groundwater level. But most importantly, unlike the philosopher, the scientist demands empirical proof, or at least *provability*. Unlike the philosopher's metaphysical theory, a scientific hypothesis needs to be in principle *testable*. Or so the cliché goes…

In the course of my four year research, I've been often struck by the gross inaccurateness of these popular iconographies, at least in relation to my research topic. Nobody can deny the enormous success of the human enterprise which is modern science, but when it comes to the study of mind and cognition, however, the caricatural contrast between the woolgathering philosopher and the attentive scientist tends to get blurry rather fast. In fact, during my research, I sometimes felt that it was precisely my role as a philosopher to keep the scientist, that is: the *cognitive* scientist, with his[2] feet on the ground. Indeed, oftentimes I couldn't suppress the – admittedly exaggerated – thought that even the wildest philosophical speculations are no match for some of the contemporary theories that are going under the banner of cognitive *science*. Yet, this reversed portrayal of the critical philosopher trying to keep the unhinged imagination of the scientist in line would be an equally flawed caricature. For on closer examination, we see that some of the most problematic concepts we find in cognitive science theory can already be found in philosophical theories long predating the birth of the so-called sciences of the mind. Perhaps, then, it would be more accurate to say that, when it comes to our inquiries into the nature of mind and cognition, the topic is remarkably non-discriminatory between the philosopher and the scientist.

[2] I do say 'his', because cognitive science is, as a matter of fact, unfortunately still a predominantly male discipline.

My philosophical research has focused on what are considered to be two of the most problematic issues within contemporary cognitive science and philosophy of mind: on the one hand, the issue of representationalist versus nonrepresentationalist approaches to mind and cognition and, on the other hand, the issue of the relation between the physical and the psychological, an issue traditionally known as the mind-body problem.

With regard to the first issue, I have in particular concentrated on the notion of internal representation, and how this notion is supposed to be doing explanatory work within cognitive science theory. The notion of internal representation forms the cornerstone of mainstream cognitive science's theoretical framework. At the same time, however, the notion is highly disputed, and – as we'll see – for good reasons.

With regard to the second issue, my investigation has centered on the functionalist notion of the multiple realizability of the mental, specifically in its relation to a potential mind-body identity theory. For despite the notion's widespread popularity in both philosophy of mind and cognitive science, multiple realization faces some serious problems of its own. As I'll argue, while both the notion of internal representation and that of multiple realization have put their stamp on our contemporary understanding of mind and cognition, they have each in their own way outstayed their welcome. Indeed, this dissertation wants to be more than an account from a neutral bystander. It wants to be a voice in the debate by contributing a set of both negative and positive arguments. Negatively, it wants to make a case against representationalist and functionalist approaches to mind and cognition. Positively, it wants to provide an argument in favor of so-called embodied, enactive, embedded and extended theories of cognition (4E-Cognition for short), as well as a defense of identity theory. More to the point, it wants to be an endorsement of a radical nonrepresentationalist E-account of the mind, combined with an identity theoretical approach to the mind-body problem.

The present dissertation consists of a collection of five separate, yet thematically intertwined papers. The first three papers are devoted to the subject of

representationalism within mainstream cognitive science, whereas the remaining two papers focus on the philosophical topic of multiple realization and mind-body identity theory. Yet, even though these papers are supposed to be sufficiently self-standing to be evaluated in their own right, the reader will nevertheless notice a certain continuity when they are read in the order in which they are presented here. This is no coincidence, for this order reflects the chronology in which they were written. The overall line of thought behind these writings is, therefore, best appreciated when they are considered in their present format.

This being said, however, the reader should be notified that at some point, the dissertation's continuity will appear to get somewhat interrupted. This topical interruption turned out to be an inevitable consequence of the ambition to investigate two different, though related subjects, i.e., representationalism within cognitive science and the mind-body problem. Mirroring this double thematic, the dissertation as a whole is divided in two main parts.

The general outline, then, is as follows: the first part consists of three papers, each laying emphasis on different problems associated with representationalism in general, and the notion of internal representation – understood as an internal (truth evaluable) content carrying entity – in particular. The first of these three papers (*Identifying Representations*) wants to focus on the ontological status of internal representations by raising the question what it means for something to be identifiable as a content-carrying representation. This is a notorious problem for representationalists, for although invoking internal representations has become a routinous affair in mainstream cognitive science theory, the question as to what it takes for some internal physical structure to qualify as a representation remains, not so much unsettled, but rather unasked.

The second paper (*Dereifying Representation*) builds further on a fundamental insight gained in the previous paper, namely that the notion of representation is essentially normative. As will be shown, crucially, this normativity is social in nature. On the reified construal of representation we find in mainstream cognitive science, however, the socio-normative character of representation is

either completely denied, or either taken to be reducible to objective causal facts. This second paper will not only argue why reifying representation is unwarranted, it will also provide an extensive explanation of where this tendency to reify an essentially normative notion comes from, and how this tendency is being maintained.

The third and final paper of the first part (*Off-line Cognition as Representation-hungry*?) deals with the widely accepted, though problematic claim that nonrepresentational approaches to cognition are necessarily limited in scope. Here, the distinction between so-called 'on-line' and 'off-line' forms of cognition will be introduced. It will be argued that the widespread assumption that nonrepresentationalist E-approaches are unqualified to deal with off-line cognition is misguided, as it is built on a confusion between the level of explanation and the level of description. Perhaps certain cognitive phenomena are best described as representational, but this does not in any way necessitate the invocation of internal representational entities for the explanation of these phenomena. Even more, it will be argued that we have good reasons to assume that it is not representations that are underlying forms of off-line cognition, but that, quite to the contrary, representation in some sense relies on certain forms of off-line cognition already being in place.

As said, the second part of the dissertation will shift its focus from the topic of representation to that of the mind-body relation. Understanding the motivation behind this thematic realignment requires some clarification. As I've indicated above, the two main subjects of this dissertation are, indeed, different, yet they are also closely related. I'll now say something about this relation.

In the literature, both the representational character of the mental, as well as its multiple realizability have been put forward as a defining feature of mind and cognition. The idea that representation is essential to mind and cognition is sometimes expressed as it being 'the mark of the mental or the cognitive'[3]. At the same time, because cognition is paradigmatically understood computationally, the alleged multiple realizability – or even, medium independence – of

[3] See, for instance, Adams & Aizawa 2008. See also Adams & Beighley 2011.

computations is supposed to underpin the claim that cognition is essentially multiply realizable or medium independent as well.[4] On this standard 'representation-plus-computation' interpretation, the relation of the mental to the body, or rather, the brain, is captured in terms of realization or implementation. Since the mental is on the representational-computational account essentially a functional category, the mental is considered to be something over and above the particular physical brain states that realize it. Moreover, the cognitive role of the rest of the body is by many still considered to be marginal at best. This, of course, does not sit well with the idea of cognition being embodied, an idea which this dissertation wants to support. In other words, if one wants to develop an encompassing argument against representationalist-computationalist accounts of cognition, one ought to be arguing against these two core commitments of standard cognitive science: on the one hand, the commitment to the idea that mind and cognition are fundamentally representation-involving; on the other hand, the commitment to the idea that the mental is essentially multiply realizable. With regard to the second commitment, the dissertation will defend a strict mind-body identity thesis, one that manages to avoid talk of realization altogether and, in addition, is fully compatible with the core tenets of the 4E-approaches to mind and cognition. The second main part of the dissertation, then, will consist of two papers. The first paper (*Multiple Realization: A Thesis with Identity Issues)* argues against the notion of multiple realization, and also against the widely accepted idea that this notion provides a superior alternative to a strict mind-body identity theory. The second paper (*Identity Reconsidered: Taking a Dual Perspective on the Hard Problem of Consciousness*) will defend a version of a strict mind-body identity theory, yet one that does not want to reduce the mental to what goes on in brains.

[4] These ideas can be found in influential work by David Chalmers and, more recently, Gualtiero Piccinini (see, for instance, Chalmers 2011 and Piccinini 2015). Chalmers emphasizes the substrate neutrality of computational processes, whereas Piccinini claims the defining feature of computation to be the medium-independent nature of the vehicles which are computationally processed.

I want to conclude this introduction with a few sentences about the dissertation's title. The main title (*Mind, Mechanism and Meaning*) will be clear enough. All three terms refer, after all, to fundamental elements of the dissertation's subject matter and are specifically meant to pick out three key ingredients co-constituting the prevalent image of the mind within mainstream cognitive science. Here, 'mind' is (still) understood as a highly complex causal mechanism performing operations over meaningful or contentful vehicles.

The subtitle (*Reclaiming Socio-normativity within Cognitive Science and Philosophy of Mind*) probably needs a little more clarification. During my four year research, I came to realize that most, if not *all* of the problems I've identified in today's dominant picture of mind and cognition, can be understood as the result of an unwarranted sidetracking, or even exclusion of the subject within scientific theory. To better explain what I mean by this, the following lines from William James might be instructive:

> By amateurs in philosophy and professionals alike, the universe is represented as a queer sort of petrified sphinx whose appeal to man consists in a monotonous challenge to his divining powers.... Reality, we naturally think, stands ready-made and complete, and our intellects supervene
> with the one simple duty of describing it as it is already. But ... [w]e *add,* both to the subject and to the predicate part of reality. (1907: 92,99)

What this dissertation wants to bring to the fore, then, is that perhaps the main reason why representationalist or computationalist approaches to mind and cognition fall short, is that it ignores the fact that "we add". More specifically, it assumes it can brush aside the intrinsic normative and subjective, or rather, intersubjective character of its basic explanatory concepts. As I will try to show, both the notion of representation, as well as that of multiple realization, rely in their own way on socio-normative practices outside of which these notions no longer make sense (or so I'll argue). In a word, a fundamental error of mainstream cognitive science lies in its assumption that it can cut loose and then – using James' phrase – 'petrify' two fundamentally normative and dynamic

entities from the socio-normative practices that constitute them. These two entities are *meaning* or *content*, on the one hand, and *classification* or *categorization*, on the other. Mainstream cognitive science thinks it can borrow these notions to be used for explanatory purposes. Unfortunately, both the notion of internal representation and that of multiple realization are – using Dennett's metaphor – built on loans it can't repay.

The representationalist needs to account for how an internal vehicle can literally be said to be carrying meaning. Within a socio-normative context, it is clear what it means for some physical object to be representing something else. But how does this work, exactly, within the physical confines of the skull?

The proponent of multiple realization is confronted with a similar problem. The multiple realization thesis contends that one and the same type of mental entity can be realized by different physical entity types. It is perfectly clear what *types* or *categories* refer to within the socio-normative practice of classifying things. But how, exactly, are we to understand these types when disconnected from these practices? For the multiple realization hypothesis to lay claim on being a scientific hypothesis for which empirical evidence can be cited, it must assume that mental state types are the kind of entities that exist independently of our normative practice of classifying the world. What, then, is the ontological status of these types? And what does it mean, exactly, to say that they can be 'realized' physically?

I'll argue that both the notion of internal representation, as well as that of multiple realization, quickly lose coherence when they are taken out of a socio-normative framework. The subtitle of this dissertation, then, is meant to emphasize that one can't simply strip certain notions from their socio-normative character and then proceed by inserting them as reified explanatory posits into a theory that is supposed to be explaining the very same cognitive capacities underlying these socio-normative entities (e.g., postulating a 'Language of Thought' to explain our linguistic/symbolic capacities). Indeed, cognitive science has been putting the cart before the horse far too long. It is time for an approach that is no longer blind to the fact that minds are always both embodied and situated in an environment and that, consequently, they are to be understood as

such. And in the human case, this environment happens to be largely socio-normatively structured. This, at least, is what I will be arguing in the following pages.

References

Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. Oxford: Blackwell-Wiley.

Adams, F., & Beighley, S. (2011). The mark of the mental. In J. Garvey (Ed.), *The Continuum Companion to the Philosophy of Mind* (54–72). London: Continuum International Publishing Group.

Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(4), 323–357.

James, W. (1907). Pragmatism and humanism. Lecture 7. In *Pragmatism: A new name for some old ways of thinking*. New York: Longman Green and Co.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

1 Identifying representations: representation as a prescriptive notion

Abstract

This paper addresses the issue of what it is for something to be identifiable as a representation. The inquiry will proceed in a number of steps. First, it will be pointed out how, in cognitive science literature, we find a tendency to conflate the idea that something can be seen as a representation with the idea that something is, as a matter of fact, identifiable as a representation. Second, it will be argued that the notion of internal representation cannot, and should not, be divorced from an understanding of what it means for an external object to qualify as a representation. In the third step, it will be argued that *all* representation involves a socio-normative or prescriptive element, and that factual identification of an object as a representation is impossible outside this prescriptive context. The conclusion argues that we should give up on the idea of discovering structures in brains that can be identified as representations.

1

Identifying representations: representation as a prescriptive notion

1.1 Introduction

In his *Representation Reconsidered*, William Ramsey notes that it has become "almost a cliché to say that the most important explanatory posit today in cognitive research is the concept of representation." (Ramsey 2007: xi) Indeed, representational posits are ubiquitous within cognitive science. Yet, this fact of representation taking pride of place within the mind's science stands in stark contrast to another, equally true cliché, namely that "there is nothing even remotely like a consensus on the nature of mental representation." (Ramsey 2007: xi) If it is true that internal representations are to be understood as theoretical entities, used to "explain observations of intelligent or adaptive behavior" (Chemero 2009: 50), one would expect that the question as to how these postulated entities are to be defined and identified has already been properly dealt with. This, however, is not the case. Regarding the question of the identification of representations, Michael Wheeler notes:

> [T]he situation is an embarrassing scandal. The idea that there are internal representations is a deep assumption of the most influential branches of philosophy of mind and cognitive science, and we really ought to know how to spot one. (Wheeler 2005: 6)

And in a recent paper, Alex Morgan and Gualtiero Piccinini claim that

> anyone seriously interested in the conceptual foundations of cognitive science must eventually grapple with what makes something a representation. (Morgan & Piccinini 2017: 11)

In the following, I will take up the question of what it means to be identifiable as an internal representation. Within cognitive science literature, the idea that something actually *is* a representation (and, therefore, that something is identifiable *as* a representation) is often conflated with the idea that something *can be seen or interpreted as* a representation. Yet, both issues are fundamentally different, and I will start by addressing this problem, i.e., the conflation of the problem of identification of internal representations with the issue of what it means to say that something can be seen as a representation. Within cognitive science literature, this conflation tends to lead authors to identify entities as representations which are in fact not to be regarded as such. Next, since there is not one well-defined notion of representation at work within cognitive science, I will have to specify what I have in mind when I use the term representation. Although there are a number of different technical notions at play within contemporary cognitive science (e.g., structural similarity based notions, action-oriented representations, distributed representations), all these notions share the core idea that internal representations are entities that, in some sense, specify "how things stand with the world" (Hutto 2009: 20) and that can therefore be said to carry a content. The next step consists in emphasizing that this central idea of representations being content-carrying entities is not something that emerged from scientific practice itself, but an assumption of cognitive science which derives from a pre-scientific source. For unlike theoretical entities like quasars or centers of gravity, in ordinary speech, we already have a notion of representation and, as it turns out, virtually all the so-called theoretical notions are dependent on a pre-scientific understanding of what it means to say that something is a representation.

Considering, first, that almost all theoretical notions of representation are modelled on a pre-theoretical understanding of what it means to say that something represents something else and, second, that a warranted theoretical use of the notion of representation actually *requires* a conceptual connection with our pre-theoretical understanding, I will address the question of what it means to

be (identifiable as) an internal representation by first explicating what it means to be (identifiable as) a more familiar external representation.

1.2 To be or not to be a representation

I'll start by delineating our main question concerning the identifiability of representations by contrasting it with the different, though often conflated idea, of being viewable or interpretable as a representation. To emphasize the importance of this distinction, I will look at a familiar discussion (the case of the Watt governor), a discussion which I believe to be largely due to a failure to properly demarcate the identification problem, leading authors to unwarranted ontological conclusions.

The Watt Governor

In his 1995 paper *What might Cognition be if not Computation?*, Tim van Gelder introduced the example of a mechanical device known as the Watt governor to unambiguously support antirepresentationalist claims. However, his representation-eschewing description of the engine's operation in terms of dynamical systems theory did, and does not convince everyone. As we shall see, even a proponent of radical embodied cognition like Anthony Chemero points out that a representationalist explanation of the Watt governor's operations remains a perfectly sound possibility. Before investigating this claim, however, let us take a closer look at the device in question.

The Watt governor is a type of centrifugal governor, a fully mechanical device designed to control the speed of an engine – in the case of the Watt governor, a steam engine – via a linear feedback control system (proportional control). Van Gelder describes it as follows:

> As the spindle turned, centrifugal force drove the balls outward and hence upwards. By a clever arrangement, this arm motion was linked directly to the

> throttle valve. The result was that as the speed of the main wheel increased, the arms raised, closing the valve and restricting the flow of steam; as the speed decreased, the arms fell, opening the valve and allowing more steam to flow. (van Gelder 1995: 349)

The Watt governor not only allowed maintaining a constant speed without system-external influence; from a historical perspective, it also helped set in motion the wheels of the First Industrial Revolution.

Van Gelder famously offers the Watt governor as an example of a dynamical system that is capable of successfully accomplishing a specific task (maintaining constant speed) in a non-computational, non-representation involving manner. Ultimately, van Gelder wants to make way for the idea "that there is in fact a currently viable alternative to the computational conception of cognition." (van Gelder 1995: 359) At the same time, however, van Gelder clearly sees the possibility of describing the device's operation as involving the use of representations because of the "initially quite attractive intuition" (van Gelder 1995: 351) to see the angle at which the arms are spinning as a representation of the current speed of the engine. However, van Gelder warns against this intuition: "[A]rm angle and engine speed are of course intimately related, but the relationship is not representational." (van Gelder 1995: 351) Van Gelder goes on to defend his antirepresentationalist claim by stressing that the arguments in its favor

> are not based on any unduly restrictive definition of the notion of representation; they go through on pretty much any reasonable characterization, based around a core idea of some state of a system which, by virtue of some general representational scheme, stands in for some further state of affairs, thereby enabling the system to behave appropriately with respect to that state of affairs. (van Gelder 1995: 351)

Van Gelder distinguishes an epistemological from a metaphysical claim[5]. The epistemological one states that, although it might be *possible* to describe the system in representational terms, compared to the dynamical explanation, it doesn't give us any additional explanatory purchase and is therefore epistemologically superfluous. Moreover, van Gelder sees in this a useful criterion for invoking representations as an explanatory posit. In general, one should ask "whether there is any explanatory utility in describing the system in representational terms." (van Gelder 1995: 352). On this point, Chemero (2009) agrees with van Gelder. However, as to the metaphysical claim, van Gelder and Chemero have diverging opinions. The former clearly thinks it's a mistake to see the Watt governor as containing representations. According to van Gelder, nothing in it is a representation, which is a claim about the nature of the system. This antirepresentationalist claim derives from the observation that mere causal correlation is insufficient for representation. Indeed, "[v]irtually everything is correlated, fortuitously or otherwise, with something else; to describe every correlation as representation is to trivialize representation." (Van Gelder 1995: 352; see also Clark 1997: 146) Representation requires something more, some "extra ingredient" (van Gelder 1995: 352), and according to van Gelder, we have no reason whatsoever to expect to find it in the Watt governor.

For Chemero, however, things aren't that simple. According to him, *given the traditional criteria*, the Watt governor can indeed be seen as a representation using device because elements in it satisfy three conditions that, according to Chemero, define representation within a traditional theory. Chemero defines representation as follows:

> A feature *Ro* of a system *S* is a *Representation for S* if and only if:
> (R1) *Ro* stands between a representation producer *P* and a representation consumer *C* that have been standardized to fit one another.

[5] Chemero 2000 and 2009 makes a similar distinction.

(R2) *R0* has as its function to adapt the representation consumer *C* to some aspect *A0* of the environment, in particular by leading *S* to behave appropriately with respect to *A0*, even when *A0* is not the case.
(R3) There are (in addition to *R0*) transformations of *R0*, *R1…Rn*, that have as their function to adapt the representation consumer *C* to corresponding transformations of *A0*, *A1…An*. (Chemero 2009: 50-51, see also Chemero 2000: 627)[6]

The Watt Governor allegedly satisfies these conditions. What we have here, according to Chemero, are so-called action-oriented representations, because the arm angles are said to "*indicate* at once the engine's speed as well as the appropriate response to that speed" (Chemero 2009: 71; m.e.). "They are both 'map and controller…*standing for* the current need to increase or decrease the speed." (Chemero 2009: 71; m.e.) The valve, then, is seen as the representation consumer, which is supposed to lead to the appropriate behavior (i.e., desired speed). And since different arm angles can be seen as standing for different engine speeds, the third condition (which states the systematicity requirement) is satisfied as well.

In his treatment of the Watt governor in a 1998 paper, philosopher William Bechtel writes the following lines, also cited in Shapiro 2011:

> The fact that the angle of the spindle arms represents the speed of the flywheel becomes more clear when we consider why it was inserted into the mechanism to begin with. …The spindle and arms where inserted so as to encode information about the speed in a format that could be used by the valve opening mechanism. (Bechtel 1998: 303)

Despite van Gelder's explicit warning against invoking representations here, Bechtel does nevertheless identify elements within the system as representations. Chemero, in his turn, argues that the centrifugal governor can be seen as

[6] As Chemero acknowledges, this definition goes back to work by Ruth Millikan. See especially Millikan 1984 & 1993.

representational and that, in addition to this, "it suggests that other dynamical systems models of cognition *can also be viewed as* having representations." (Chemero 2009: 71; m.e.) He concludes that antirepresentationalists better give up on trying to argue against representationalists on a metaphysical level. This seems to be an over-hasty conclusion.[7]

1.3 Is/As

As my emphases in the above quotations already indicate, I want to draw the reader's attention to a tendency we find in Chemero, but also in other texts dealing with the nature of representation. On closer inspection, Chemero's understanding of the metaphysical claim seems to be sheltering two different claims. On the one hand, the metaphysical claim is presented in terms of something's *being or not being* a representation: 'First, one might be making a claim about the nature of cognitive systems, namely that nothing in them *is* a representation. ….' (Chemero 2009: 67; m. e.) This claim, in other words, concerns our present issue of the identification of representations. On the other hand, however, Chemero's suggestion of giving up on antirepresentationalist metaphysics finds its justification in the idea that, given his definition, some entities *can be seen as* representations. For instance:

> Since van Gelder offers the Watt governor as a prototypical dynamical system and a new paradigm for the modeling of cognition, the fact *that it can be seen as representational* is significant, and it suggests that other dynamical systems models of cognition *can also be viewed as* having representations. (Chemero 2009: 71; m. e.)

Lawrence Shapiro, in discussing connectionism, conflates things similarly, though perhaps more unambiguously: "Connectionist networks…represent: they enter states *that can be interpreted as* being about features of the world" (Shapiro 2011: 115; m. e.) Crucially, however, saying that something *is* a representation or

[7] For further discussion on the Watt Governor, particularly from a radical enactive perspective, see Hutto & Myin 2013: 59-62.

that something actually represents, and saying that something *can be seen, viewed or interpreted as* a representation or as representing is saying two very different things. It is perfectly acceptable to say that something can be seen as something else, without it *being* that something else. In terms of our central question regarding the identification of representations: it is perfectly possible to be viewable as a representation, without therefore being identifiable as a representation. However, when dealing with the notion of representation, apparently, this rather trivial, though crucial distinction tends to get overlooked.[8] Since Chemero's metaphysical claim – which he thinks the antirepresentationalist should abandon – is obviously concerned with the ontological issue whether or not some entity can actually be identified as a representation, he should be sufficiently clear on whether he takes his definition of representation to support the idea that some elements within the Watt governor can be properly identified as representations, or whether it allows one to only see or interpret them as such. If only the latter were true, there would be insufficient grounds to give up on antirepresentationalist metaphysics. For the relevant question is not whether or not something *can be seen as* a representation; what needs to be affirmed is that something *should* be seen as a representation, because of its *being* a representation.[9] Similarly, Nico Orlandi stresses the importance of distinguishing 'as if' claims from actual identity claims. With regard to the widespread idea that the visual apparatus is an inferential mechanism, she writes:

> Many things act *as if* they perform inferences. In this sense, the visual apparatus would be inferential just because pretty much everything is, which is a fairly uninteresting thesis. We need to understand the claim more robustly… (Orlandi 2014: 18)

[8] The distinction is also noted, though not elaborated, by Martin Flament-Fultot: "One may assert or deny that some object *is* a representation. But one may also assert or deny that some object is *better seen as* a representation." (M. Flament-Fultot 2014: 150)

[9] As we'll see, in case of representation, quite paradoxically, something's *being* a representation depends on the socio-normative fact that something *should* be seen as something else. Here, the 'ought' determines the 'is'. This is one of the central claims of this paper, which will be elaborated in the following sections.

And with regard to the idea that connectionist networks contain semantic content, she rightly notes:

> Again, ascription of semantic content to connection strengths or clusters is both superfluous and arbitrary, if it is based solely on the consideration that the connection *can be seen that way*. (Orlandi 2014: 101; m.e.)

Compare, for example, the difference between claiming that hearts can be seen *as if* they are pumps on the one hand, and that hearts *are* pumps on the other. Affirming the first kind of claim usually comes down to acknowledging that, because of some common features, an object resembles 'x' in some respect; however, whether or not it can also be identified as an instance of 'x' is at this point still undecided; the latter requires affirmation on whether or not the similarity is *relevant* for proper identification as classification.[10] So, for example, with functional notions such as pumps, the relevant similarity does not lie in the object's physical features, but in the sameness of the purpose they serve, and the way this is being accomplished – in this case, mechanically moving fluids around in a system. What's important here is that the question of identifying 'x' has logical-conceptual priority over the question of whether something can be seen as (an) 'x'. It makes no sense to say that something can be seen as (an) 'x' if we do not already know what it means to actually be (an instance of) 'x'. In short, both the claim that something can be seen as a representation and the claim that something actually is a representation presuppose an understanding of what it means to be properly identifiable as a representation. The present paper wants to provide a contribution to that understanding.

[10] 'Proper' meaning here: in accordance to some accepted set of classificatory criteria. In chapter four, much more attention will be devoted to the topic of identification as classification (which will be referred to as C-identification), as well as the subject of the classificatory criterion.

1.4 Representation deflationism

Before continuing, however, an anticipatory remark must be made about what is nowadays referred to as representation deflationism (see Hutto & Myin, forthcoming; see Ramsey forthcoming). With regard to the ontological status of representations, next to internal representation realists and eliminativists, a number of authors are now trying to occupy a third, kind of midway position. In the above, I have been emphasizing the importance of distinguishing between identificatory claims (something *is* a representation) and, what we could call, interpretatory claims (something can be seen *as* a representation). Representation deflationism can be characterized as denying the importance of this distinction. The deflationist holds that, if a system can be usefully (i.e., predictively) viewed as containing internal representations, then there are representations in the system. In one sense, then, on a deflationist view, there is no 'conflation' between a system containing things which can be viewed as representations, and a system's actually containing (internal) representations. If we can usefully see it as a representation, it *is* a representation.

Deflationism faces a number of serious issues. Hutto & Myin note:

> A fully deflationist theory of mental representation has to pull off a neat trick. On the one hand, it must abandon any commitment to the existence of bothersome properties associated with mental representations – properties that happen to be canonically associated with them. Yet, on the other hand, it must retain the idea that mental representations feature in our best characterization of what lies at the basis of cognition. (Hutto & Myin forthcoming)

These authors identify a number of critical shortcomings with different deflationist theories (see also Ramsey forthcoming). However, for my present purposes, it is unnecessary to go into their analysis. The important thing to note here is that the question of the identifiability of representations remains an important issue for the representation deflationist as well. If representation deflationism is the view that something *is* a representation if it can be usefully

viewed as a representation within the context of scientific explanation and prediction, the deflationist still owes us an account of how to identify internal representations. For even if we allow the deflationist logic that something is an 'x' if we can usefully see it as an 'x'; in other words, even if we agree to coalesce the truth conditions of our ontological claims about what really exists with the success of our predictions, we still need to know what it means to be identifiable as an 'x' in order to know what it is to be 'viewable' as an 'x'. As already argued above, it is a matter of conceptual logic that we can't view or interpret something as an 'x' without first having some idea of how to identify an 'x'. The deflationist idea that something is an 'x' if we can usefully see it as an 'x' still relies on an understanding of what it means to be identifiable as an 'x'. Deflationism does not offer us any story here, it merely begs the question.

1.5 Theoretical and Pre-theoretical Representations

The point of the previous part has been to show how, within philosophy of cognitive science, the idea of being a representation tends to get conflated with the idea of being 'viewable as' a representation. And I have argued that the latter is not sufficient for actually being identifiable as a representation. This, of course, has not brought us any closer to an answer to the question of what it means to be identifiable as an internal representation. In this regard, William Ramsey's investigations in *Representation Reconsidered* (Ramsey 2007) provide a more promising perspective. As we've already seen, Ramsey holds that any theory that posits internal representations has to meet the challenge of showing that these entities are actually doing something "recognizably representational in nature" (Ramsey 2007: 28), meaning that the theoretical notion of a representation should be sufficiently similar to our pre-theoretical notion of what it means to represent something else. This idea can already be found in Dennett (1978):

> Whatever *mental* representations are, they must be understood by analogy to *non*mental representations, such as words, sentences, maps, graphs, pictures, charts, statues, telegrams, etc. (Dennett 1978: 175)

Call this Dennett & Ramsey's Prescriptive Claim. Unfortunately, our pre-theoretical understanding of what it means to be a representation does not allow for easy conclusions. There's considerable variety in the things we functionally classify as representations (linguistic entities, models, maps, traffic signs, pictures, drawings, art-work…). At the same time, we see that almost[11] all technical notions of internal representation at work within cognitive science are in fact modelled on one of these more ordinary, external representations. Depending on the theory, internal representations are conceived of, sometimes as linguistic entities, as internal pictures of some kind, as maps, or as models – indeed, representations are commonly modelled on models (see Godfrey-Smith 2004). Yet, although there are many different kinds of representational theories, according to Ramsey,

> all share the core assumption that mental processes involve content-bearing internal states and that a correct accounting of those processes must invoke structures that serve to stand for something else. (Ramsey 2007: xi)[12]

Considering the observation that virtually all technical notions of internal representation are based on a pre-theoretical familiarity with external representations; considering also Dennett & Ramsey's Prescriptive Claim that there *should* be this connection between pre-theoretical and theoretical notions; considering these two elements, then, warrants the idea that we can gain more insight in what it means to be identifiable as an internal representation by first trying to deal with the question of what it means to be identifiable as an external representation. This idea would be unwarranted if cognitive scientists would actually be availing themselves to their own technical notions of representation, ones that can be sufficiently divorced from our ordinary usage of the term. But, as things presently stand, this is simply not the case.

[11] I say 'almost', for it is not immediately clear how the notion of distributed representation, as it is used within connectionist theories, can be said to be still related to an ordinary public representation. However, if Ramsey is right, it is also unclear how these notions can still be counted as *representational* notions. Indeed, much of Ramsey's work (Ramsey 2007) is aimed at showing why invoking representations within connectionist theories is unwarranted.

[12] This formulation rightly highlights the fact that 'representation', internal or not, is a functional notion.

1.6 Representation as a Prescriptive Notion

As I will argue, a crucial step towards a better understanding of what it means to be identifiable as a representation lies in the acknowledgement that representation *always* involves a specific cognitive capacity, namely the capacity to relate to, or take up an attitude towards something absent through something that is present. Saying of something that it stands for something else comes down to saying that we should use our capacity to see it as something else which is, in some sense, absent. However, something can be absent in more than one way. To better understand representation's close connection to the absent, it is worth to briefly distinguish these different, what I will call, *modes of absence* here. As we'll see, to these four modes of absence correspond four categories of public representations.

Four modes of absence:

First, something can be absent in a spatiotemporal sense, when it is not here right now, or not anymore, or not yet.

Second, something can be absent in the sense of not present in the 'right' way, as when something is too big, too small, too fragmented, too fast, too complex…to relate to directly. Third, something can be absent in the sense of being non-existent in the way fictions are. Fourth, something can be absent in the sense of being abstract (mathematical objects, functions, meanings, values…).

Humans typically have the ability to relate to all of these (differently absent) things by means of something that is present, i.e., a representation. Representation always involves such an act of relating to the absent through the present. To illustrate this point, consider these four paradigmatic examples of representation that correspond to the four modes of absence:

1.We relate to a deceased person via a picture (*spatiotemporal absence*),

2. We relate to the solar system via a model (*structural absence*),

3. We relate to unicorns via a drawing (*fictional absence*),

4. We relate to justice via a statue of a blindfolded woman holding a scale and a sword (*abstract absence*).

The importance of our capacity to relate to the absent can hardly be overestimated. Without it, all forms of sign or symbol involving behavior, including of course linguistic behavior, would be impossible. We use this capacity when we decide to take an umbrella with us after seeing grey clouds appearing on the horizon, and we also use it when we see shapes on a paper as text. A full analysis of this fundamental cognitive capacity falls, however, well beyond the scope of this paper. What I do want to elaborate on here is the observation that exercising the ability to relate to something as something else is, in itself, not enough for that something to be properly identifiable as representing that something else. Indeed, virtually everything can be seen as a representation because *we can* (in the sense of 'having the capacity to') see virtually everything as standing for something else.[13] But, as I've argued extensively above, being viewable as an 'x' does not make it genuinely identifiable as an 'x'. In addition to our capacity to see something as standing for something else, a socio-normative or prescriptive element is required. For something to actually *be* a representation, it isn't enough that we *can* relate to something as something it is not; it is a requirement that we *should* relate to it in this way. This is why I insist on thinking of representation as a prescriptive notion. In the end, representation is a socio-normative phenomenon, and it is only against a background of social normativity that certain objects can become, as a matter of fact, identifiable as representations. This, of course, needs some clarifying. I'll start by giving some examples that should help elucidate what I have in mind when I say that the identification of representations always involves a prescriptive element.

1.6.1 Ought determines is: first example

Suppose you are in your local library, working at one of the desks. Suddenly, your phone starts ringing in your jacket pocket. As you reach for your phone to switch it off, the reader in front of you clears his throat. Probably some of us have been in a similar situation. To be clear, I do not mean having been in a situation where

[13] Nelson Goodman writes: "[A]lmost anything may stand for almost anything else." (Goodman 1968: 5)

someone's throat-clearing is used as an expression of annoyment, mild indignation, or as a non-verbal instruction to alter the other's socially unacceptable behavior or whatever. I take it we have *all* been 'throat-cleared' or 'shushed' at some point. What I *do* mean is having been in a situation where we don't know for sure whether someone is clearing his throat because he wants to express something, or because he is really just clearing his throat. It is not just the fact that we don't know how to interpret the hawing sound, or that we are wondering what the sound means ('Am I annoying him?' or 'Does he have something in his throat?' or 'Is he sick?'). The problem is that we haven't got enough to go on to decide whether it is *meant to mean something* at all. Put in terms of representation, it is not the case that we don't know *what* the sound represents, but rather, that we don't know whether or not it represents at all. And this latter possibility crucially depends on whether or not the sound is *supposed* to represent something. For the sound to actually *be* a representation, it is a necessary condition that it *should* be seen – or, rather, heard – as something it is not. Assuming that, in our example, the throat clearing is indeed supposed to serve as a kind of social sanction, the sound can indeed be properly identified and described as a representation. Crucially, however, *the descriptive fact is preceded by a prescriptive fact*: the sound does indeed stand for something else, but only because it is *supposed* to stand for something else. In other words, here, 'ought' determines 'is'.

1.6.2 Ought determines is: second example

I want to look at a second example of what I hold to be essentially the same phenomenon, but which is more in line with our discussion. It's a variation on thought experiments by Putnam (1981: 1), French (2003: 1473), and Ramsey (2007: 23). Suppose you're stranded on an uninhabited beach. You come across a strangely arranged structure of driftwood, which on closer inspection appears to be spelling out the English word GOD[14]. Now, even if you felt absolutely sure to

[14] Ramsey's example is slightly different in that his driftwood spells out 'UNINHABITED BEACH' . In French's example, we are asked to imagine the wind and sea having carved the Lorentz transformations

be the only one on the island, seeing the driftwood in this configuration would still have a specific effect on everyone with sufficient reading skills. First, we would automatically see it as a word, rather than 'just' driftwood; that is, it occurs naturally that we read it. *Not* seeing it as something that is supposed to be read requires effort. Second, we would all experience a kind of interruption or disruption within our, what I would call, attitudinal continuity, which is very similar, or indeed perhaps identical, to the experience we have when it is still an undecided matter whether someone's throat clearing is the expression of mild irritation or simply the physiological result of having something in one's throat.[15] On intersections like these, we don't immediately know what to do in the sense of not knowing what we are *supposed* to do. We don't know what stance or attitude we're *supposed* to take up, because we don't know whether our capacity to relate to the absent is being called upon or not.

With these examples in mind, I think we can come up with an answer to the question what it means to be identifiable as a representation, and it is a deceptively simple one: Saying of an object that it is a representation is nothing else than the explicit affirmation that we *should* make use of our cognitive capacity to relate to it as something else that is, in some sense, absent. At its core, representation is therefore not to be understood as a descriptive notion referring to some naturally occurring state of affairs, but first of all as a prescriptive one. Much of the confusion surrounding the notion of internal representation – but also surrounding the debate about scientific representation[16] – stems from conflating a descriptive factuality, namely the obtaining of a similarity relation, with a prescriptive element coming from the socio-normative practices within which something can only first become identifiable as a representation. The problem of coming up with some acceptable set of necessary and sufficient

in the sand of a beach. Putnam on his part asks us to imagine ants tracing lines in the sand that look like Winston Churchill. To make my example at least minimally realistic, I opt for a slightly more probable, though also sand-involving scenario.

[15] In his famous paper *Freedom and Resentment*, Strawson distinguishes between the objective attitude and participant reactive attitudes. I believe the experience I'm referring to in my example is closely related– if not identical to – the experience of the tension Strawson detects between a participant reactive attitude on the one hand, and the objective attitude on the other.

[16] See for instance Suarez 2003, Callender & Cohen 2006, Chakravartty 2009.

conditions for representation, which still occupies numerous scholars in discussions about both mental and scientific representation, has its source in the erroneous assumption that representation is essentially a name by which we denote some mind-independent, as well as social-norm-independent object. The fact that we often *do* use the term for naming a specific set of physical objects (the set including linguistic symbols, pictures, drawings, maps, scientific representations…) should be understood as deriving from the fact that it is a socio-normative matter that we *ought* to relate to these objects in a certain way, that is, in a way that requires drawing from our cognitive capacity to relate to the absent. I'll illustrate my point with a third and final example.

1.6.3 Ought determines is: third example

Consider Ramsey's observation that some washed up driftwood on a shore that happens to be "arranged in a way that maps a course to a nearby lake" (Ramsey 2007: 23) does not intuitively count as a representation. Now, according to Ramsey, it does become a representation when we actually *use* it as standing for a part of the terrain. Again, when it comes to accounting for representation, I think putting the emphasis on the functionality of structural similarities is a mistake. What is overlooked here is the difference between saying that something *can be used as* a map (because of some structural similarity relation) and saying that something *is* a map (and therefore identifiable as a representation). The reason we hesitate to label something that happens to be usable as a map 'a representation' is because we know of these things that we aren't supposed to relate to them in a way we are supposed to relate to actual maps. And, vice versa, it is also this prescriptive nature of representations that explains why a map that is insufficiently similar to the terrain and is therefore unusable, can still be identified as a representation. An example of the latter would be a map that does not, or no longer, sufficiently resemble the territory. An example of the former would be the abovementioned accidental arrangement of driftwood: The accidental structural resemblance guarantees that seeing the parts of this structure as standing for parts of the island will pay off. But that doesn't make it a

representation. Accidental arrangements of driftwood simply aren't *supposed* to represent anything; that is, we aren't supposed to take up, what I would call, a representational attitude towards these and other phenomena. The socio-normative fact that we *are* supposed to take up this attitude towards certain objects (including those devices known as scientific representations) is precisely what allows these objects to be genuinely identifiable as representations. Indeed, to properly understand the factual nature of representations, we have to reject the erroneous assumption that, 'a genuine fact must be a matter of the way things are in themselves, utterly independently of us' (McDowell 1998: 254). Putting matters somewhat sloganesque: when it comes to *being* a representation, 'ought' determines 'is'.

Before returning to the subject of internal representations, I want to point out that the observations above can also account for why none of the things that can genuinely be identified as representations are naturally occurring entities. As a rule, representations are not only agent-dependent, but *person*-dependent entities. But to this, we should immediately add a cultural-anthropological observation: which entities are considered to be person-dependent might vary considerably, depending on the socio-normative context in which they are functioning. Consider the already mentioned example of clouds. Although we can see grey clouds as an indication of rain, in no intuitive sense do rainclouds qualify as representations. We could explain the fact that we know that it will rain because of the clouds in terms of reliable covariance (cf. van Gelder's interpretation of the Watt Governor), but we wouldn't say that the clouds actually represent, or stand for the fact that it is going to rain. They don't, because they are not *supposed* to stand for something. We *would*, on the other hand, say of the weatherman's pictures of rain clouds that they are actually representing the possibility that it might rain tomorrow. Unlike actual rain clouds, these entities *are* supposed to say something about the way the world is (or is going to be). More precisely, unlike clouds, *we* are *supposed* to relate to these entities as something else (the predictive proposition 'that it is going to rain').

But now imagine a culture in which clouds, or some other celestial phenomenon like solar eclipses or rainbows, are believed to be a kind of medium by which some supreme being communicates with the community of devotees. Take, as an actual example, this passage from Genesis: “God said, "This is the sign of the covenant which I am making between Me and you and every living creature that is with you, for all successive generations; I set My bow in the cloud, and it shall be for a sign of a covenant between Me and the earth."” (Genesis 9: 12-17) Within the socio-normative context in which rainbows are supposed to be related to as standing for the covenant between God and all living creatures, rainbows actually are identifiable as genuine representations. But the fact that they should be related to as standing for something else is itself dependent on the belief that rainbows are person-dependent entities (where, in this example, the person is God). In itself, naturally occurring entities, or better: entities that are believed to be naturally occurring, never qualify as representations. It is only in virtue of their position within a socio-normative context that natural entities can acquire the status of a content-carrying representation. Beyond this context, however, there is no further fact of the matter that can help us decide whether or not something is to be identified as a representation, because beyond this context, there is no ‘ought’.

1.7 Identifying internal representations

What does all of this mean for cognitive science’s notion of internal representation? I have argued that representation *should* be understood in a certain way. It is important that we qualify this claim. When one says that some notion should be understood in this or that way, one might be saying two different things. First, one might be simply stating that some word is to be understood as defined by the author; its content is then simply given by stipulation. Second, however, one might also be making a claim about how some phenomenon of which we already have some understanding, can be understood

in a more insightful manner.[17] In the course of this paper, I have tried to make clear why my claim should be understood in this second way. The above account tries to contribute to a better, more comprehensive understanding of a complex phenomenon – representation –with which we all have some pre-theoretical familiarity, and on which virtually all technical notions of internal representation are modelled. Provided, then, that the story above manages to account for what it is to be identifiable as a representation, we are now in a position to better assess the idea itself of there being such entities in our brains, which we can, in principle, detect and identify. And to be clear, within cognitive science, this idea is very much alive. In fact, according to at least one researcher, it is already an established fact: not only do we really *have* representations in our brains, we are also able to detect and identify them. In an online interview, renowned behavioral neuroscientist Kate Jefferey claims that the existence of internal representations

> has become apparent in the last few decades , because we've been able to look into the brain, with our electrodes and so on, and we can see that there clearly is an operation of internal representations.[18]

In light of the above, it is hard to see how this claim makes sense, at least not if Jefferey's notion of representation wants to be not entirely divorced from all existing notions of representation. For if there is one thing that the above account has tried to bring to light, it is that the notion of representation has no place within the context of neuroscientific discovery. Representations just aren't the kind of things that can be seen (Jefferey), spotted (Wheeler), detected, or otherwise discovered by observing brains. It is simply a mistake to think that a both mind-dependent *and* socio-normative phenomenon can be objectified and

[17] Philosopher Filip Buekens makes this important qualification with regard to the notion of truth. Theorists – and philosophers in particular – tend to appropriate existing notions, often rendering them virtually unrecognizable. Buekens quotes Foucault on truth: ""Truth" is to be understood as a system of ordered procedures for the prediction, regulation, distribution, circulation and operation of statements." (Foucault 1972: 132, cited in Buekens 2014: 26) As Buekens notes, this is not what we ordinarily mean by 'truth' and is, therefore, in an important sense *not* how we should understand it.
[18] See http://philosophybites.com/2016/12/kate-jeffery-on-concepts-and-representation.html

inserted into the context of scientific explanation. The idea that some physical structure can be identified as a representation only makes sense in relation to certain socio-normative practices, in which our capacity to see something as something it is not is called upon. Granted, the claim that representations aren't the kind of things that can be discovered needs to be somewhat modified. Discovering representations is indeed in some sense possible, but these discoveries are the province of the cultural anthropologist, not the brain scientist. Again, outside a socio-normative context, the idea that something is in and by itself identifiable as a representation is misguided. There are no intrinsic representations, where 'intrinsic' is supposed to capture the idea that an object can be said to be a representation in virtue of its (non-normative) properties. This means that, from the disengaged, non-normative perspective of the scientist (for instance, the neuroscientist), representations simply fail to show up, regardless whether they are internal or external.

I want to conclude this section with a comparison of my account with the anti-representationalist position advocated by REC (Radical Embodied/Enactive Cognition). In their critical *Radicalizing Enactivism*, Daniel Hutto and Erik Myin confront the representationalist with, what they call, the Hard Problem of Content. Roughly put, their claim is that anyone who wants to invoke content-bearing structures like internal representations as naturalistically credible explanatory posits owes us a naturalistic explanation of content. In other words, cognitive science theorists are free to use contentful representations in their explanations of cognition, provided they have a story to tell about how these entities acquire their truth evaluable content. Since there is no such story on the market, invoking internal representations to do explanatory work is unwarranted. And since there does not seem to be any plausible theory of content on the horizon, cognitive science should perhaps better give up on the idea of explaining cognitive phenomena in terms of internal content-bearing structures. Hutto and Myin's line of argument, although persuasive in my opinion, is sometimes met with the reply that cognitive science's reliance on internal representations is not

at all unwarranted, even though we have as yet no accepted explanation of how internal structures can have content. Invoking representations is very much warranted – so the argument goes – by the "stunning successes"[19] of representationalist theories. Furthermore, it is argued that, although we do not have a scientific theory of how internal entities come to acquire content, this does not mean that science will never figure it out. In any case, the fact that science hasn't figured it out yet should not stop scientists from using internal representations as hypothetical entities to do explanatory work.[20]

In light of the above, it should be clear that I can only agree with REC that cognitive science theorists should give up on the idea of there being physical structures in the brain that can be identified as representations. In an important respect, however, my approach differs from Hutto and Myin's argument. Their emphasis is placed on the representationalist's inability to *explain* internal representation. I've tried to approach things differently, not in terms of *explanation*, but in terms of *explication*. Indeed, before we can even sensibly raise the question as to how internal representation could be explained, we should start by asking what it would even mean for something to be (identifiable as) an internal representation. This latter question is an explicatory, not an explanatory one. And the explication above has hopefully made it clear that internal representations aren't the kind of entities that need to be explained; they are the kind of entities that need to be explicated away.

1.8 Concluding remarks

To cognitivists, the above account of representation must seem far removed from what they have in mind when they invoke representations as an explanatory posit. Of course, cognitive science theorists are free to come up with stipulative definitions of representation, but then they should be clear on the definitions' relationship to the actually existing phenomenon. If, on the other hand, the theorist wants to model his or her notion of internal representation on actual

[19] Shapiro 2014: 214.

[20] We find this line of argument in the work of Michael Rescorla. See, for instance, Rescorla 2016: 13.

instances of the kind of representation with which we already have some familiarity – which is clearly the case – then it should be made clear how the idea of there being naturally occurring content carrying entities in the brain can be rendered compatible with the above account. My current view is that it can't. As a kind of summary, I want to run through three of the main arguments that motivate my position. First, representations do not so much explain cognition, but are themselves dependent on a particular cognitive capacity, itself in need of an explanation, namely the capacity to relate to the absent through something which is present. As said, this (typically human) cognitive ability requires much more investigation, which falls beyond the scope of this paper. It does seem clear, however, that such an investigation will involve, among other things, examining the workings of the imagination. Second, representation is not only a mind-dependent phenomenon, it also always depends on, and is embedded in, a socio-normative context. It is here we find the basis for its objective character, and this basis is, as I've argued, prescriptive in nature. Objects can only be said to be representations if we *ought to* relate to them as something else which is absent. It would be a mistake to think we can make sense of this requirement within the context of scientific discovery. The brain, or any other supposedly cognitive system, simply does, and cannot provide the right kind of normativity which representation requires. From the objective perspective by which the scientist hopes to discover explanatorily relevant entities, the phenomenon of representation simply disappears. Third, in our present account of representation, the spatial opposition internal/external becomes irrelevant. On my account, representation is to be understood in terms of socially regulated cognitive behavior. Adding to this the spatial qualification of being internal or external is irrelevant. Indeed, we do call some spatially localizable objects representations (so-called public representations). Yet, as I've argued, the fact that we do this should be understood as both deriving from, and depending on our attitudinal behavior and our ways of socially regulating and consolidating these practices. Representations are never simply 'in there', just as they are never simply 'out there'.

References

Bechtel, W. (1998). Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cognitive Science 22*: 295-318.

Buekens, F. (2014). *De transparantie van waarheid*. Acco: Leuven.

Callender, C. & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria, 21*: 67–84

Chakravartty, A. (2010). Informational versus functional theories of scientific representation. *Synthese 172*: 197-213.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*: 625-647.

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

Clark, A. (1997). *Being There*: *Putting Brain, Body and World together again*. MIT Press: Cambridge, MA.

Dennett, D.C. (1978). *Brainstorms*: *Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Foucault, M. (1972/1980). Truth and power: an interview with Michel Foucault. In C. Gordon (Ed.), *Power/Knowledge: Selected Interviews & Other Writings 1972-1977*. New York: Pantheon Books.

Flament-Fultot, M. (2014). On genic representations. *Biological Theory 9*: 149–162.

French, S. (2003). A model-theoretic account of representation (or, I don't know much about art…but I know it involves isomorphism). *Philosophy of Science 70*: 1472–1483.

Godfrey-Smith, P. (2014). Signs and symbolic behavior. *Biological Theory 9*: 78-88.

Goodman, N. (1968/1976): *Languages of Art: An Approach to a Theory of Symbols*. 2nd ed., Indianapolis: Hackett.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, and D. Rumelhart (Eds.), *Philosophy and Connectionist Theory*. Hillsdale, N.J.: Erlbaum.

Hutto, D. D. (2009). Mental representation and consciousness. In W. P. Banks (Ed.), *Encyclopedia of Consciousness 2*: 19-32. Oxford: Elsevier.

Hutto, D.D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content.*Cambridge, MA.: MIT Press.

Hutto, D.D. & Myin, E. (forthcoming). Deflating deflationism about mental representation. In J. Smortchkova, K. Dolega and T. Schlicht (Eds), *What are Mental Representations?* Oxford University Press.

McDowell, J.H. (1998). *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA.: MIT Press.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, MA.: MIT Press.

Morgan, A and Piccinini, G. (2017). Towards a Cognitive Neuroscience of Intentionality. *Minds and Machines*. Doi. 10.1007/s11023-017-9437-2

Orlandi, N. (2014). *The Innocent Eye*: *Why Vision is not a Cognitive Process*. New York: Oxford University Press.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Ramsey, W. M. (2007). *Representation Reconsidered.* New York: Cambridge University Press.

Ramsey, W.M. (forthcoming). Defending representation realism. In J. Smortchkova, K. Dolega and T. Schlicht (Eds), *What are Mental Representations?* Oxford University Press.

Rescorla, M. (2016). Bayesian Sensorimotor Psychology. *Mind & Language 31*: 3–36.

Shapiro, L. (2011). *Embodied Cognition.* London: Routledge.

Shapiro, L. (2014). *Radicalizing Enactivism: Basic Minds without Content*, by Daniel D. Hutto and Erik Myin (Review). *Mind* 123:213–220.

Strawson, P. F. (2008). Freedom and Resentment and Other Essays. London: Routledge.

Suarez, M. (2003). Scientific representation: against similarity and isomorphism. International Studies in the Philosophy of Science 17: 225-244.

van Gelder, T. (1995). What might cognition be, if not computation? Journal of Philosophy 92: 345-381.

Wheeler, M. (2005). Reconstructing the Cognitive World. Cambridge, MA: MIT Press.

2 Dereifying Representation

Abstract

The notion of internal representation is without doubt one of the most central explanatory posits within mainstream cognitive science. At the same time, however, there is much controversy about how, exactly, we should conceive of internal representations. This is unsurprising, considering the fact that there are multiple notions of representation at work within cognitive science. This paper focusses on the classic, and still prominent construal of representation, on which the notion is understood as an internal, semantically evaluable content-carrying entity which can in principle be individuated and identified. I will refer to this conception as the reified notion of representation. On the reified construal, representations are conceived of as thing-like entities with thing-like properties. The paper consists of two main parts. In the first part, I want to critically assess the tenability of this reified notion of representation. I will argue that the notion is ultimately untenable as it is built on a confusion between the descriptive and the prescriptive. Representation is a socio-normative notion which loses its meaning outside a socio-normative context. In the second part, I will investigate what motivated, and continues to motivate cognitive science theorists to adopt this incoherent conception of representation as an explanatory posit. I will distinguish five such motivations. As I'll show, crucially, these motivations do not derive from scientific practice itself, but from certain contingencies we find outside of this practice. This further presses the issue of whether the reified conception of representation is warranted as an explanatory posit.

2 Dereifying Representation

2.1 Introduction: a generic notion of representation

Within mainstream cognitive science, internal representations take pride of place within the theorist's explanatory apparatus. For many, providing explanations of cognitive phenomena just *is* providing an internal representation invoking story. Yet, despite their centrality to cognitive science, when it comes to the exact nature of these representations, most theorists remain silent. Truth is that, within representationalist cognitive science theory, representations aren't so much argued for, but rather assumed. As one author puts it: "Representations are invoked even *before* the theory starts." (François Tonneau: 2011/2012) On the other hand, however, some philosophers and theorists *have* gone to considerable lengths when it comes to explicating and elucidating the notion of internal representation (for more recent accounts, see, for instance, Mark Rowlands 2006, William Ramsey 2007, Anthony Chemero 2009, Nico Orlandi 2014). Yet, no two accounts of representation provide us with the exact same characterizations. This is unsurprising, given the fact that there simply is not one notion of representation at work within cognitive science. As Ramsey notes:

> [I]n the case of representation, there is actually a cluster of very distinct notions that appear in very distinct theories. (Ramsey 2007: 4)

However, when it comes to mainstream computational/informational accounts, a generic notion of representation can nevertheless be discerned. Following Frances Egan's phrase, on 'the Standard View' (Egan 2012), representations are understood as internal, content-carrying physical entities over which computations are defined, and which have an explanatory role within scientifically respectable theories of mind and cognition. It is this influential notion of representation which figures prominently in classic work by Newell and Simon, Fodor, Pylyshyn[21] and others and it is this notion which, to this day,

[21] See Newell and Simon (1976), Fodor (1975, 1980, 1981), Pylyshyn (1980, 1984).

remains a conceptual cornerstone for mainstream cognitive science, even if it isn't always recognized as such. It is this notion I'll be focusing on here. For reasons explained below, I will be referring to this notion as the *reified* notion of representation.

2.2 The reified notion of representation

As Egan (2012) points out, on the Standard View, representations are thought of as intrinsically dual in nature. This duality is usually captured in terms of the vehicle/content distinction. From one point of view, internal representations are conceived of as physical entities. This is considered a requirement if representations are to be at once internal, as well as causally efficacious. And, in addition, having a physical nature is also taken by many as synonymous with being naturalistically acceptable[22]. For many, providing a naturalistic account *just is* providing an account which is principally reducible to a physical account. On the other hand, next to their being physical (cognitivists of a functionalist stripe prefer 'physically realized', which of course doesn't make them any less physical in their particularity), representations are also conceived of as content-carrying. Accounting for this essentially metaphorical[23] claim is a different story altogether and much philosophical labor has gone into the subject of naturalizing content. Below, this subject will be treated in more detail. For now, however, it is first of all important that we get clear on what – within mainstream cognitive science – the 'content' metaphor is supposed to pick out exactly. Michael Rescorla explains that the notion of content needs to be cashed out in terms of semantic evaluability, which is indeed a widely held – though often implicit – view within cognitive science. Rescorla writes that the content of an internal representation

[22] 'Naturalistically credible', meaning here 'compatible with the natural sciences, and especially physics'. If representations are thought to be causally efficacious and internal – and therefore spatiotemporal – entities, they cannot be non-physical.

[23] Below, the nature of this metaphor will be discussed in more detail.

> represents the world as being a certain way. We can ask whether the world is indeed that way. These states are semantically evaluable with respect to such properties as truth, accuracy, and fulfillment. (Rescorla 2016: 17)

Moreover, according to Rescorla (2016), the details about how, in the end, content is to be cashed out empirically shouldn't worry us too much. What is important is that "the science itself" [24] makes clear that the idea of content-carrying representations is indispensable within explanations of cognition.

But despite what "the science" might have to say about internal representations, it is undeniable that these stipulated theoretical entities are modeled on a kind of ordinary public representation, that is, a kind of linguistic representation; more specifically, they are modelled on assertions or claims or judgments or other expressions in relation to which it makes sense to ask whether what is being expressed "is indeed that way". As is by now clear, there are serious difficulties with thinking about representations as internal entities with assertoric content. One of the problems is that it directly comes up against what Daniel Hutto & Erik Myin have dubbed the 'Hard Problem of Content ' (Hutto & Myin 2013; see also Hutto & Myin 2017). These authors point out the fact that we have as yet no naturalistically credible way of making sense of there literally being semantic content/information in the brain, and that, in addition, we have no reason whatsoever to assume that the issue is going to be resolved any time soon. As Mason Cash puts it, "a naturalistic reduction of content facts to physical facts" remains an "unfulfilled desideratum" (Cash 2009: 134). And although at least one champion of internal representation claims that this problem has already been solved years ago by teleosemantic accounts (Marcin Milkowski 2015), this kind of optimism seems to miss the full scope of the difficulties. In the end, teleosemantic theories are still faced with the problem of showing how an appeal to biological mechanisms sensitive to covariance explains truth evaluable

[24] As Hutto & Myin note, Rescorla "is quite fond of reporting what the science says or does, and of deploying "the science" to "rebut various views"… Rescorla (2016) uses the "science says/does" illocution no less than fifteen times in a single article, reporting what the science: describes (pp. 4, 27, 28); posits (p. 17); remains neutral about (p. 20); explains (p. 22); illuminates (p. 22); cites (p. 23); assigns (p. 23); may not vindicate (p. 25); does not employ (p. 27); does not advance (p. 28); aims to provide (p. 29); purports (p. 29); and extends (p. 31)." (Hutto & Myin 2017: 182)

content. As Hutto & Myin point out, "covariation doesn't entail or constitute content" (Hutto & Myin 2013: xvi). In other words, teleosemantics still owes us an account of how information-as-covariance (which is ubiquitous) can provide the explanatory grounds on which to base a theory of the kind of semantic information representationalists like Rescorla – and many others – appeal to when they invoke a semantically evaluable notion of content.[25] And although the latter doesn't share the same unrestrained optimism of Milkowski, his suggestion is that the fact that we don't have a theory of mental content *yet* shouldn't stop us from invoking representations, as long as appealing to these content-carrying entities brings with it further "stunning successes"[26] for cognitive science. As he puts it: "[O]ne may legitimately describe mental activity at an abstract level that prescinds from neural implementation details" (Rescorla 2016: 13).

However, the problem is that, for Rescorla's claim to be *itself* legitimate, we must at the very least assume that the terms used to "describe mental activity at an abstract level" are themselves coherent. What I want to show is that the reified notion of representation as employed by cognitive science theorists to describe *and* explain mental activity at an abstract level is, indeed, incoherent. In the following, I will argue that the idea of internal semantically evaluable content-carrying entities makes little sense, as it is ultimately grounded in a category mistake, namely the mistake of conceiving of an essentially prescriptive notion as a descriptive one. More specifically, I will argue that the idea of a content – carrying representation, internal or not, is a socio-normative notion and has therefore no place in causal accounts of cognition, at least not on the *explanans* side of the account. Nevertheless, the reified and descriptive conception of representation remains indeed standard within mainstream theories of mind and cognition, most distinctly within computationalist approaches. If my argument is sound, it follows, then, that to the extent that these theories rely on the reified notion of representation, they are flawed.

[25] For a critical evaluation of the idea that teleosemantics can provide a naturalistic account of semantically evaluable content, see Hutto & Myin (2013), especially chapter 4.
[26] Shapiro 2014: 214.

2.3 Representation as a socio-normative notion

I want to start my argument by reemphasizing that, in general, technical notions of representation conceptually rely on our ordinary pre-theoretical understanding of representations, and that, in particular, the idea of internal truth evaluable representations derives from our familiarity with those 'things' we call assertions (or claims, or statements, or judgments…). The first thing to note is that, in case of an ordinary assertion like 'the cat is on the mat', its being truth evaluable is not just one of its properties. An assertion is not in any ordinary sense a thing with properties. An assertion essentially just *is* truth evaluability. Another way of saying this is that the notion of truth evaluability exhausts the notion of an assertion. So anything that is truth evaluable can be said to be an assertion, and vice versa. We might proceed by asking what it is that makes some 'thing' an assertion. To prevent us from starting off on the wrong foot, answering this question should begin with allowing the possibility that it might not be a *what* that makes things into assertions, but a *who*. As some of the most prominent philosophers of language – including Wittgenstein, Dewey, Quine and Davidson – have been at pains to show, language is essentially social. And whether we are dealing with verbal utterances, written sentences, sign language, braille, or any other semantically evaluable physical entity, these 'things' only have truth conditions if they are *supposed* to have truth conditions. Quoting Davidson on this matter:

> An utterance has certain truth conditions only if the speaker intends it to be interpreted as having those truth conditions. (Davidson 2005: 50)

The fact that certain objects can be said to be carrying truth evaluable content (and, therefore, be identifiable as assertions) is, in other words, a socio-normative fact. It is only because it is *supposed* to stand for something else (say, an assertion), that a physical structure (a written sentence, say) becomes properly identifiable as a representation. We naturally fail to notice this because hearing or seeing or touching (in case of braille) words and sentences *as* words and sentences has become so much part of our cognitive nature as to make us think of these entities as "perfectly ordinary particulars" (Shea: 2013, see further). This

is to treat something which is not a particular as a particular, hence an act of reification. This reified thinking – conceiving of something as thing-like, or as an "ordinary particular" – about truth evaluable representations, however, is a misconception. This misconception explains why all naturalizing attempts of reducing content to a phenomenon explicable in an objective physical or biological vocabulary have failed. They were bound to from the outset. With regard to intentionality in general, Mason Cash makes the point that "nothing intrinsically *has* intentional states" (Cash 2008: 101). Likewise, I want to argue that nothing intrinsically *is* a content-carrying representation. As with intentional states, the reason is that, crucially, contentful representations are *constitutively* socio-normative, both with regard to their 'vehicle' aspect, as well as with regard to their content. In the following, I'll try to explain what both claims mean exactly.

2.3.1 Representations and normativity

The intuition that assertoric content (i.e. content that says how things stand such that things could be different) is in some sense normative is widely shared amongst philosophers. It has been endorsed in some form by, for instance Wittgenstein, Sellars and Kripke, and it is also being espoused more recently by philosophers like Robert Brandom, Paul Boghossian, Allan Gibbard, Michael Luntley, Daniel Whiting and Mason Cash (to name just a few). But despite their shared recognition of the normativity of content, there is little agreement on how this normativity is to be understood exactly. Michael Luntley, for instance, claims:

> Our everyday notion of content is a normative concept. The content of a thought, or the content expressed by a sentence, is characterized in terms of the circumstances under which it is correct or incorrect to assent to it. It is expression of content that makes our sentences subject to evaluation for truth or falsity. The terms of semantic evaluation (truth or falsity) are normative concepts. If a sentence is false, then, other things being equal, you ought not to utter it. (Luntley 1999: 2)

Luntley's emphasis on truth or correctness conditions is already an important indicator of the normativity of content, or better – and correcting Luntley – of *assertoric* content. Yet it nonetheless insufficiently grasps the specific socio-normative sense authors like Robert Brandom and Mason Cash have in mind when they hold that content is *constitutively* socio-normative in nature. This kind of normativity is indeed best expressed by 'ought', but it is wider than Luntley's 'ought' in 'you ought not to utter a false sentence'. There is much more going on here, a fact that is very concisely captured by Wilfrid Sellars' remark that content is "fraught with ought" (Sellars 1962: 44). This compact phrase manages to highlight two crucial elements about the normativity of content: on the one hand, the fact that normativity is involved in more than one way (it is *fraught* with ought); on the other hand, the fact that the normativity involved is *social* normativity (it is fraught with *ought*). Furthermore, and perhaps more importantly, when it comes to physical representations (public linguistic items or perhaps neural structures in the head), social normativity is involved *constitutively*, both with regard to what is doing the representing (the vehicle) as with regard to what is being represented (assertoric content). Below, I'll explain what this means separately.

2.3.2 Socio-normativity with regard to representational vehicles

The claim that representational vehicles are constituted by 'oughts' can be paraphrased as the claim that some physical structure is only properly identifiable as a representation because of the constitutive involvement of social norms. To better see what this means, I'll use an example[27]. A man is cheating on his wife. Whenever she's out of town, he switches on only one of two porch lights, so his mistress would know it's safe to come over. The point is that, because the mistress knows she's *supposed* to see the single porch light as meaning something like 'my wife is away', the object actually acquires the status of a truth evaluable object. The object has been endowed with truth or

[27] The example I'm using is actually a variation on an illustration by David Lewis, recently picked up again by Peter Godfrey-Smith. See Lewis (1969) and Godfrey-Smith (2013).

correctness conditions, which is to say it has been *constituted* as an assertion, not only for the promiscuous couple, but in principle for anyone else who knows about the lovers' secret arrangement. It has now become a fact that the husband ought not switch on just one porch light if his wife is still in the house, because it has become a (social) fact that 'one porch light on' now means 'wife not at home'. Conversely, if the husband wants to express the fact that his wife is in the house, he ought not switch on just one light. But the fact of the porch light being a contentful representation with correctness conditions is itself constituted by the 'oughts' and 'ought nots'. Without these socio-normative constraints, the physical object could never have acquired its status as an assertoric representation. The descriptive fact of the object being properly identifiable as an assertoric representation follows from prescriptive socio-normative constraints. But there is absolutely nothing about the porch light itself (its properties) that would allow us to attribute it with content. For suppose that the husband and his wife are both at home, but for some reason, and unbeknownst to the husband, in the evening his wife forgets to turn off the second porch light. Drama ensues. But what has happened here is strictly speaking not the result of the porch light being interpreted wrongly; the mistake actually lies in the mistress interpreting something which is not *supposed* to be interpreted to begin with. In case of the wife leaving one light on, the mistress mistakenly believes the object to be semantically evaluable or interpretable, which it isn't. In other words, she *ought not* have attributed the porch light with semantically evaluable content in case of the wife leaving it on. This shows that the identifiability of the object as an assertoric representation can never be reducible to the object's properties. It is, after all, exactly the same object in both scenario's.

The example can be broadened to *all* representations with assertoric content. Whatever physical structures count as a representation with truth evaluable content (whether it's porch lights, sound waves, scribblings or small convex dots on a piece of paper, hand gestures or neural structures), this constitutively depends on what we *ought* to count as a representation with such content. This is *always* to some extent an intersubjective matter of convention, whether it is the

local *ad hoc* convention between a man and his mistress (the porch light) or the global on-going conventions between a whole community of language users (linguistic items). Of course, the question itself as to *what* we ought to count as a truth evaluable representation and what not remains to be answered (for instance, when can a child's utterance 'mama' be actually identified as a proper assertion, containing for instance the truth evaluable content 'you are my mother'?), but the important point for now is that the answer will be one in terms of 'oughts'. Ultimately, the fact that certain physical structures are socio-normatively constituted as assertoric representations cannot be understood independently of our practices of content attributions. And, as we'll see next, the latter too is a socio-normative affair.

2.3.3 Socio-normativity with regard to representational content

As indicated above, in this section, I want to investigate the claim that not only the vehicles of assertoric content, but the truth evaluable contents themselves are constituted socio-normatively as well. For this purpose, I will draw heavily from the work of Mason Cash, who I think presents the most convincing story here. He defends the thesis that

> the meanings of people's utterances, the contents of their thoughts, and thus the informational contents of neurological representations that are said to implement these thoughts, are all constituted by the normative social practice of ascribing and accepting intentional states as reasons for actions. (Cash 2009: 134)

In a number of papers, Cash – himself influenced by the work of Brandom – presses the point that the kind of norms constitutive of content attributions are *intersubjective* norms. He distinguishes these norms from two related, though different kinds of norms. According to Cash, intersubjective norms are not reducible to 'objective correctness conditions' (Boghossian 2003, 2005; see also Luntley: 1999), nor can they be identified with 'subjective norms of rationality'

(Gibbard: 2003, 2005)[28]. Rather, content is constituted by 'intersubjective norms of rationality' (Cash 2008: 99). The key to understanding the intersubjective account of content lies in the recognition that content does not exist outside the socio-normative practice of giving and asking for reasons for our assertions, as well as our actions. Crucially,

> [t]he ability to give and ask for reasons for ascriptions of intentional states to one another, then, depends on an understanding of socially specified norms regarding when one is entitled to ascribe a particular intentional state to an agent. These norms make appeal to what the agent has perceived, as well as what they have said and done, and also to other intentional states the observer has previously been entitled to ascribe to that agent. (Cash 2008: 99)

The "particular intentional state" we are interested in here are assertoric beliefs with truth evaluable content. Applied to beliefs, then, what Cash is saying is that one can only ascribe a particular truth evaluable content to an agent when one is entitled to do so. This entitlement on behalf of the observer can be put in terms of 'oughts': you ought only ascribe a certain content to an agent if that agent has the ability to provide the relevant reasons for entertaining that content. This means that there are also 'oughts' applying to the agent: the agent ought to be able to provide relevant reasons for his or her contentful assertions. Put otherwise, the agent has to be able to take up *responsibility* for his or her assertions, the latter being a matter of giving reasons for holding a certain content. But this is an entirely intersubjective matter. Both the practice of giving reasons, as well as the kind of reasons that are supposed to be given or asked for, are socio-normative practices root and branch.

The above can be made more clear by means of another example. Suppose someone at your office murmurs the words: 'The building is on fire.' For you to

[28] 'Oughts' understood as reducible to 'objective correctness conditions' (Boghossian) implies that one ought to entertain a content only if that content is true. As Boghossian explains it, "for any p: one ought to believe that p only if p." (2005: 210) In other words, the objective truth of p determines the normative correctness of the content of the belief that p. Gibbard's account of 'subjective norms of rationality', on the other hand, conceives of these 'oughts' as arising from 'rules of rationality'. Entertaining a belief that p not only implies knowledge of objective correctness conditions under which p is true. It also implies that one (subjectively) entertains other beliefs, desires, intentions… which are related to p through certain rules of rationality.

ascribe to him the truth evaluable content of the building being on fire at that moment, your colleague ought to be able to give you the relevant reasons for entertaining that assertion. This means, for one, that he has to be able to understand under what conditions the assertion would be true. If your colleague would utter these words every time the printer is jammed, you would not be entitled to ascribe to him the content 'the building is on fire'. For you to be entitled to ascribe this content to him, he has to be in principle able to give you some reasons for why he entertains that content. Crucially, however, for these reasons to be *the right* reasons, they must be intersubjectively shared. For someone's reasons of entertaining a certain content to be acceptable, someone else must be in principle entitled to hold the same content for the same reasons.

It is useful to contrast the example of the colleague making claims about the building being on fire with that of a fire alarm. Some theorists hold that there is a sense in which a triggered fire alarm represents the truth evaluable content 'the building is on fire'. Just as we can ascribe contentful assertions to competent language users, we can legitimately ascribe content to such devices. They carry information about how things stand with the world. This alleged fact is then usually explained in terms of causal covariance. From this point, then, we are only one step away from applying the same kind of explanation of content to living organisms, and in particular, their brains. Unlike the actual fire detector, the biological devices that tell animals (including humans) that there's a fire are not manufactured in electronics factories, they have been naturally selected for. Yet, just as the actual sounding fire alarm carries semantically evaluable content about the state of the environment, so too do our own naturally evolved *internal* fire detecting mechanisms.

It should be clear that people who advance a normativity thesis of content discard both such causal and teleosemantic accounts of content. As Hutto and Myin argue, causal covariance, whether we find it in artificial devices or evolutionarily evolved systems, does not constitute content. (see Hutto & Myin 2013: 67) In addition, I would argue that the idea that certain devices like fire detectors or thermometers can be, in and of themselves, ascribed semantically

evaluable content derives from a kind of conceptual error which might be characterized in terms of a metonymy or displacement. Rather than attributing content to the agent for whom the fire alarm (or the porch light) *is supposed to mean something*, content is ascribed to the physical object which is closely associated with it (a fire detector, a porch light, ink marks on a piece of paper, acoustic vocal patterns etc.). Returning to our example of our co-worker claiming that the building is on fire, we see that a sounding fire alarm is not itself identical to, or even analogous with an agent making contentful assertions. Rather, the sounding alarm is *one of the relevant reasons* our colleague 'ought' to give for entertaining the content 'the building is on fire'. But this, in turn, is always also a socio-normative fact in that it is only within a certain socio-normative context that sounding fire alarms ought to be interpreted that way (just like a single burning porch light is only interpretable as 'wife is away' within a certain socio-normative context). An animal might perhaps learn to see the connection between a sounding fire alarm and an actual fire, but on the socio-normative account, we cannot invoke any content here because there is no legitimate reason for invoking any socially shared 'oughts'. However, if we only ought to ascribe content to a creature that ought to be able to give the right kind of reasons for entertaining that content, then it seems that we ought not ascribe contents to animals or little children (and certainly not to inanimate objects like fire detectors, thermometers or, indeed, computers). This brings us to the first of a number of objections against the intersubjective account of content and representation.

2.3.4 First objection: What about animals and infants?

If content ought only be ascribed to agents participating in the sphere of socio-normativity, does this mean that non-participants like animals and small children cannot possess contentful thoughts or beliefs about the way the world is? According to Cash,

> [t]here is a crucial difference between having an intentional state, in the sense that an observer can ascribe it to you, and having the concept of an intentional state, in the sense that you can also ascribe it to yourself...Having the ability to think about thoughts is a prerequisite for being able to not just *have* a belief, but *to self-ascribe* a belief, and to understand it *as a belief*. This is uniquely the domain of socialized, language-using persons. (Cash 2008: 101)

It seems undeniable that when we observe the behavior of certain animals and little children, it often makes perfect sense to ascribe intentional states to them. Crucially, however, we do so only because we recognize similarities between our own intentional behavior and the actions or movements of animals and infants. Yet, it is precisely through these practices of intentional state ascription that an infant, but not an animal, learns to develop the relevant concepts necessary for the self-ascription of contentful states. As Cash puts it:

> Animals and infants can have intentional states, in the sense that we can sensibly ascribe them. We can describe an infant reaching for a bottle as 'wanting to eat'. We can sensibly describe a lion running after a gazelle as trying to catch and eat that gazelle. But they don't have intentional states in the self-aware way we fully socialized adults do. (Cash 2008: 101-102)

Through socialization, the child eventually becomes a participant in the practice of giving and asking for reasons. In time, we can no longer only ascribe intentional states to the child, but we can ascribe to the child both *self-ascriptions* of intentional states, as well as intentional state ascriptions to others. By comparison, we might attribute intentional states to our pets on a daily basis, yet they will never develop the cognitive capacities to self-ascribe a belief, nor to ascribe a belief to us. Perhaps certain animals can be said to have socio-normative practices of their own, but the specific linguistic practice of giving and asking for reasons, as a necessary condition for having a belief *qua* belief, appears to be an exclusively human affair.

Although it is unclear whether Cash is aware of this, it should be mentioned that the above ideas concerning the attribution of intentional states to children can

already be found quite literally in the work of Lev Vygotsky[29]. Anyone familiar with Vygotskyan psychology will recognize in Cash's notion of 'socialization' the Vygotskyan concept of 'internalization'. In David Bakhurst's discussion of a 1931 passage from Vygotsky[30], the parallel in thought between Mason Cash and the early 20th century Soviet psychologist is unmistakable:

> Vygotsky proposes the following account. He argues that the symbolic relation between noise or gesture and object or action is set up only in the context of the child's relations with other people. From birth, the child participates in situations in which his or her behaviour is significant for others. The child's movements (or utterances) are attributed meaning by the surrounding adults. However, though the adults treat some of the child's movements as signs, the sign does not enter the situation as something that has meaning for the child: It acquires its significance only in virtue of the function the movement takes on for the adult participants. This function is then said to undergo a process of "internalization" in which the movement (utterance) through which it is realized becomes endowed with meaning *for the child.* (Bakhurst 1991: 76-77)

The passage is followed by an excerpt from Vygotsky's *The History of the Development of the Higher Mental Functions* (1931/1978) in which Vygotsky makes this part of his theory clear by means of an example, namely the development of the pointing gesture. According to Vygotsky, in a first stage, pointing starts out as merely a movement infants happen to make when reaching to grasp something. When they stretch their arms and fingers to try and grasp an object, the child is doing something very similar to what adults do when they point at an object, at least superficially. The next stage crucially requires the involvement of another person:

> When the mother comes to the child's aid and realizes his movement indicates something, the situation changes fundamentally. Pointing becomes a gesture for others. The child's unsuccessful attempt engenders a reaction not from the object *but from another person.* Consequently, the primary meaning of that unsuccessful grasping movement is established by others. Only later, when the child can link

[29] Cash makes no reference to Vygotsky in the publications I'm aware of.
[30] I am grateful to Karim Zahidi for drawing my attention to this passage.

> his unsuccessful grasping movement to the objective situation as a whole, does he begin to understand this movement as pointing. At this juncture there occurs a change in that movement's function: From an object-oriented movement it becomes a movement aimed at another person, a means of establishing relations. *The grasping movement changes to the act of pointing.* (Vygotsky 1931: 143-144/1978: 56; quoted in Bakhurst 1991: 77)

To sum up, then, on Cash's socio-normative account, animals and infants cannot be said to have the capacity to ascribe intentional states to others, nor themselves, and in that sense cannot be said to *have,* e.g. beliefs. Only through a process of socialization, which crucially involves intentional-state ascriptions from the adult to the infant, can the child gradually develop into a competent participant within the socio-normative practice of giving and asking for reasons.[31]

2.3.5 Second objection: Isn't the intersubjective account viciously circular?

The objection has been raised[32] that, because the norms that constitute intentional states are themselves dependent upon intentional states, the account is viciously circular and could never get off the ground. Although some theorists are quick to dismiss an account at the first signs of circularity, Cash usefully reminds us of the difference between vicious and non-vicious circularity:

[31] Still in relation to the first objection, Gallagher and Miyahara claim that socio-normative accounts (which they refer to as neo-pragmatist accounts), "…run into a different problem, namely, in their attempt to account for our commonsense ability to recognize intentionality in the behavior of a variety of non- or pre-social entities (…) According to neo-pragmatism, something is an intentional agent only if it acts according to norms that are socially based. (…) More generally, if a creature (e.g., a non-human animal) completely lacks understanding of social norms, and is not expected to act in accordance with such norms, it seems that the ascribing of intentionality itself would be inappropriate. And yet we do ascribe intentionality to animals, and others who lack understanding of social norms (e.g., presocial infants). (…) Neo-pragmatists, then, seemingly fail to explain our everyday practices of ascribing intentionality to such creatures." (Gallagher & Miyahara 2012: 124-125) I admit to not fully understand the relevance of these authors' considerations here. Their critique seems misplaced as it is unclear, first, how Gallagher and Miyahara reach their conclusion that neo-pragmatists "fail to explain our everyday practices of ascribing intentionality" based on the considerations leading up to their conclusion; second, why explaining ascriptions of intentionality to non- or pre-social creatures would be a special problem, specifically for *socio-normative* accounts. Moreover, as we've seen, Cash *does* provide a possible answer in terms of our recognition of the similarities in movements between animals/infants and adults. Gallagher and Miyahara, however, claim that "[w]hat this proposal entails is not clear."(Gallagher & Miyahara 2012: 125) It is not clear to me, however, why the authors find Cash's proposal "not clear".

[32] See, for instance, Kalish 2005: 246 and Hattiangadi 2006: 235.

> The apparent circularity of explaining intentional states as constituted by the norms of a practice that is itself constituted by the intentional states of those who enforce and follow those norms, however, need not be a barrier to seeing intentionality as a naturalistic phenomenon[33]. This, I will show, is a non-vicious circularity. (Cash 2008: 111)

Cash deals with the alleged circularity or infinite regress by means of an analogy.

> Compare the apparent circularity of explaining the existence of a particular human being, by appeal to the prior existence of that person's parents, also human beings, and so on back through their ancestry. This is not a viciously circular explanation. If we trace the ancestors back further and further in time, we see organisms that are progressively less and less human-like, eventually going back to primate ancestors, to the first mammals, and to the first multi-cellular organisms, and so on. (Cash 2008: 112)

According to Cash, evolutionary accounts such as these are not viciously circular at all: "We can show this apparently vicious circularity to be a non-vicious, recursive, circularity." (Cash 2008: 111). Cash is not all that explicit about it, but it is important that we are clear about what this difference means. Calling an account viciously circular is an epistemic claim. It is a claim about our understanding of things, not about the things themselves. It makes no sense to say of the objects under investigation that they are viciously circular. By contrast, however, saying that an account is (non-viciously) circular might very well be a claim about the nature of the studied objects or phenomena. The account would then be circular in the sense of being an account *about* circular (or periodical) phenomena. Now, the gradual evolutionary process that eventually leads up to the existence of creatures with the right abilities to self-ascribe contentful beliefs is ontogenetically best thought of as a (highly complex) circular process. It is only through generations and generations of intersubjective pattern-repeating that something like a socio-normative context can begin to emerge. Rather than being an impediment, these recursions of intersubjective behavior are precisely a necessary ingredient of a naturalistic explanation of intentional states.

[33] This use of the notion of 'naturalistic' is reminiscent of Hutto & Satne's so-called 'relaxed naturalism'. See Hutto & Satne (2015).

Ultimately, we could say that it's precisely the repeating patterns that constitute, or rather, *enact*, the specific socio-normative practices in which ascriptions of belief can only begin to make sense. Spelling out the details of these patterns should therefore be an integral part of any naturalistic[34] theory of content.

2.3.6 Third objection: How can content still play a causal role?

The problem being raised here can be put as follows: on the one hand, we want the specific content of an intentional state to play an explanatory role in our theories of behavior. We want to say that it is John's believing that the building is on fire that causes him to shout out 'Fire!'. On the other hand, however, the intersubjective account conceives of contents as socio-normative entities that are irreducible to physical properties. Content facts do not reduce to physical facts. How, then, can such entities be causally efficacious? Our treatment of this third objection will be a bit more lengthy, as the question is particularly relevant for our discussion. The problem brings out an essential contrast between the intersubjective normative account of representation and the still prevailing reified picture we find in cognitive science. In fact, when we ask ourselves how the reified picture is motivated, causal efficaciousness seems to be precisely one of the main reasons for assuming representations must be 'thing-like', or at least reducible to 'thing-like' entities (presumably neural structures). For many, for content to be part of a naturalistically respectable explanation of behavior, it simply *must be* reducible to something physical. Jerry Fodor, for instance, expresses this as follows:

> I want a naturalized theory of meaning: a theory that articulates in non-semantic non-intentional terms, sufficient conditions for one bit of the world to be about (to express, represent, or be true of) another bit. (Fodor 1987: 98)

Despite what Fodor wants, on the intersubjective normative account, the assumption that 'bits of the world can be about other bits of the world' in and of

[34] Again, see Hutto & Satne (2015) (see also the previous footnote).

themselves comes out as incoherent.[35] Nevertheless, it is this assumption that keeps spurring researchers to look for the sufficient conditions of 'aboutness' or content in brains. If cognitive science wants to make any progress on the subject of meaning, it should relinquish the hard-wired assumption that there is a sense in which neural states intrinsically have contents, and that these contents are what they are in virtue of the obtaining of certain observable, discoverable, measurable, and ultimately re-identifiable physical facts. Again, on a socio-normative account, no parts of the world, including neural parts, *intrinsically* have content. But what is perhaps equally important: even if we somehow *could* make sense of the idea that brain states possess content in and of themselves, this still does not answer our question as to how content *qua* content can be causally efficacious. In other words, the problem of the causal role of content is not only a problem for the socio-normative account, it is just as much an issue for those endorsing the reified notion of internal representation. How, then, do 'socio-normativists' deal with the problem of content and causality?

Cash's take on the problem is based on a distinction by John Haugeland. In one of his essays[36], Haugeland introduces the distinction between phenomena that are *instituted*, and those that are *constituted* by a normative practice. A promise would be an example of a phenomenon that is normatively *instituted*, rather than constituted. As Cash explains:

> What is and is not a promise is entirely up to us. If we no longer gave and accepted promises then there would be no promises. Furthermore, promises and their causal powers place no constraints on the practice of promising. A promise has no causal powers in itself. There are no facts about promises that apply independently of the practice of promising. (Cash 2008: 109)

[35] Similarly, Hutto writes: "I find the very idea that parts of the world or parts of organisms might *be* content-involving simply incoherent." (Hutto 2008: 57, emphasis in the original).
[36] Haugeland's distinction is actually itself inspired by work of John Rawls and John Searle. See Haugeland 1998: 318.

Things like murder weapons, on the other hand, are said to be normatively *constituted*[37]. Contrary to a promise, a murder weapon (a knife, say) is an object that has "an independent existence, and there are facts about its causal powers." (Cash 2008: 109) But, of course, the fact that the object is properly identifiable *as a murder weapon* is normatively constituted by our normative practices in which certain actions, but not others, count as murder[38]. In itself, however, there is no such thing as murder, and so neither can there be said to be murder weapons. Cash's point is that contentful representations belong in this category of normatively constituted, rather than instituted, phenomena. When we ascribe certain contentful states to agents as the cause of their behavior, there is always some physical entity involved that we identify as the relevant behavior, whether it is a bodily movement, a vocal utterance or a brain mechanism. These are entities with objectively causal powers. But they are not entities that can be said to have a certain content in any objective sense, that is, in and of themselves. If they do acquire intentional status, this can only be because these physical entities have become subjected to our normative content attributions. In other words, the fact that certain causally efficacious physical entities can be identified as contentful is a normative fact coming from our content ascriptions, not from the physical properties of these entities considered in isolation from our content ascriptions. To most neuroscientists, this will sound rather counterintuitive. They might hold that the normativity thesis must have gotten things backwards. On the intersubjective normative account, explaining content starts with social

[37] Contrary to Cash, I would hold that promises are not that different from baseballs, and that they are just as much normatively *constituted*. Cash's claim that "a promise has no causal powers in itself", whereas objects like baseballs do, seems misguided. The point is that, just as with baseballs, there are no "promises in itself". What we have are certain acoustic vocal patterns, or scribblings on a piece of paper, that, like any other physical entity, do have causal powers, just as the spherical leathery object we call a baseball has causal powers. And just as with normatively constituted objects like baseballs, allegedly normatively instituted objects like promises put constraints on our normative practices as well. Just as not any old object can be used as a ball within the game of baseball, so too not just any acoustic utterance or written shapes can be used within the 'game' of giving and accepting promises. Saussure's arbitrariness of the signifier only goes so far. Within a given linguistic community, it really does matter what physical means are being used to make a promise.

[38] The example is reminiscent of Hume's famous discussion of, what is now called, the naturalistic fallacy. In the first section of part one of book three of his *Treatise*, Hume writes: "Take any action allow'd to be vicious: Wilful murder, for instance. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call vice. In which-ever way you take it, you find only certain passions, motives, volitions and thoughts. There is no other matter of fact in the case. The vice entirely escapes you, as long as you consider the object." (Hume 2002/1739-40)

practices involving first of all normative judgments about observable bodily behavior (actions). Only secondarily does the account consider potentially relevant brain mechanisms. To the neuroscientist, this must feel like driving against the traffic. A proper neuroscientific explanation of intentional behavior should start by looking at brain mechanisms with assumed content, and then see how these mechanisms causally lead up to the observable bodily behavior which can then be objectively described by others as the result of a certain content carrying brain process causing the behavior. It can, however, be shown that it really is the brain scientist, and not the philosopher who is getting things backwards. This becomes clear when we consider the question of why a brain scientist would start looking for certain neural mechanisms to explain an intentional state – a belief, say – in the first place. It is not that neural processes wear their alleged content – and, therefore, their identity as a belief or a desire or whatever – on their sleeve. The very idea that certain brain events might be identifiable as a belief depends entirely on our reasons for ascribing a belief to that person, and these reasons do not come from the scientific practice, but from the socio-normative practice of ascribing content to certain physical events, but not to others. When a person bumps her head, there is, of course, certain brain-activity as well. But no neuroscientist would in this case assume that this neural activity is identifiable as a belief. The scientist knows, just like everybody else, that truth evaluable beliefs aren't the kind of thing we would ascribe to a person at the time when she's bumping her head. Moreover, suppose there is some random sentence going through her head that might well be caused by the bumping. For instance, when she bumps her head, for some reason, the sentence 'cows are not yellow' is 'triggered'. This would still not be a case of having a belief. The reason is that having a belief requires a proper normative context in which we can in principle give relevant reasons for holding the belief. This means that we must be able to give at least some explanation of how the belief hangs together with other beliefs. But we cannot do this for a randomly triggered sentence like 'cows are not yellow', so we wouldn't be entitled to properly ascribe this belief to the person bumping her head. And neither would she be entitled to ascribe it to herself.

To sum up: it might be the case that one day, neuroscience will discover a kind of neural activity that is causally involved whenever a person can be said to have a belief. But this latter fact of the agent being ascribable with a belief remains socio-normatively constituted and does not exist independent of a socio-normative background. Thinking that investigating all the causal properties of a neural state will one day explain why it is identifiable as a belief is like thinking that examining all the causal properties of a knife will eventually explain why it is a murder weapon. As Hume already knew, thinking that this can be done is committing, what Ryle later called, a category error.

2.4 Ought determines is

The thesis that social normativity sometimes *constitutes* new entities, such as semantically evaluable representations, might be hard to accept, not only for neuroscientists, but for anyone endorsing the belief that "the physical facts fix all the facts"[39]. Nevertheless, it seems that Hume's guillotine does not cut both ways. Perhaps you can't get an 'ought' from an 'is', but the other way round seems much less barred. This is actually more obvious than it sounds when we take into account that human beings have never been individuating or classifying the world exclusively, or even primarily, based on its objective physical properties. The most basic way of structuring the world is not in terms of what things 'really' are in themselves – from an objective, scientific perspective –, but what they are for a subject, and first of all, for a group of subjects. Basic classificatory acts involve, in other words, an element of significance.[40] And in some cases, we ourselves endow objects with a significance which they wouldn't have based on their physical (causal) properties. In case of basic representations like markings

[39] Rosenberg 2011: 26

[40] Even the most mad dog naturalist does not escape this element of significance. For even if one claims that all there is are bosons and fermions and that everything can be explained in terms of these elementary particles and their properties, the fact that there is a distinction between an explanandum and an explanans to begin with relies on a structuring activity that can't itself be objectively motivated in terms of bosons or fermions, but that refers back to the differentiating subject, for whom something might appear as something that needs explaining.

on bones[41], for instance, their significance lies both in the fact that they can be interpreted, and that they can be interpreted as meaning this or that (e.g. one notch stands for one piece of livestock). Yet the fact that they *can* be interpreted results from the fact that they *ought* to be interpreted. This requires, on the one hand, the ability to take up a specific evaluative attitude towards an object, namely one that refrains from evaluating the object with regard to its potentially salient physical properties. From a cognitive science perspective, we could say that the required cognitive skills need already be in place. On the other hand, it requires some minimal form of social convention that first of all establishes the normative truth of the object having truth conditions, namely those conditions that make it true that some object is interpretable or not. These conditions are socio-normative. So unless we can make sense of the idea that there are socio-normative conditions at work within the brain, we have no reason to expect we can make sense of cognitive science's reified picture in which there exist internal content-carrying entities that, like truth evaluable assertions, represent how the world is, such that it could be otherwise.

2.5 Whence reification?

In the following sections, I want to make an attempt at identifying (at least some of the) implicit conceptual factors or assumptions that motivate, reiterate and reinforce the reified picture of representation. As we shall see, I will especially be focusing on background assumptions that are not themselves motivated by any scientific research or empirical data, and which I shall refer to as 'contingencies'; what they have in common is that they are all contingent in the sense that they are not so much grounded in empirical research or scientific practices, but rather in pre-scientific ways of speaking and, consequently, thinking. I shall distinguish and discuss five of these contingencies:

[41] I'm specifically thinking of the set of notched bones found at Border Cave in South Africa, dating to approximately 44,000 BCE, which were in all likeliness used for the purpose of counting.

1. Hypostatization

2. Ambiguity regarding 'representation'

3. Ambiguity regarding 'states'

4. Property-talk

5. Vehicle/content distinction

Before discussing these 'reifying contingencies', a quick word on the idea of reification itself. It has recently come to my attention that the notion of reification is not always being used in the same sense. In a guest-lecture[42], Michael Rescorla uses the notion of reification as synonymous with making something quantifiable, assuming that this is the original sense in which Quine used the notion. Furthermore, Rescorla thinks science is warranted to reify representations in this sense because we need to make representations countable in order to theorize about them. I think Rescorla is mistaken, both in believing that reification equates to countability and in assuming that this is also what Quine had in mind. Reification, as the etymology of the term clearly indicates, is the practice of conceiving of something as a *thing* (the Latin 'res' means 'thing'). In German, we get 'Verdinglichung' ('Ding' meaning 'thing'), in French we get 'chosisme' ('chose' meaning 'thing'), and in English, we also find some authors using 'thingification[43]' as an alternative to reification. Furthermore, this is also how Fodor and Pylyshyn use the notion in there influential 1981 paper on Gibson:

> Having introduced the (purely relational) notion of states of affairs *containing information about* one another (i.e. being correlated) Gibson then slips over into talking of *the information in* a state of affairs. And, having once allowed himself to *reify* information in this way (*to treat it as a thing, rather than a relation*), it is a short step to thinking of detecting the information in the light on the model of, for example, detecting the frequency *of* the light…(Fodor & Pylyshyn 1981: 167; added italics mine)

[42] Rescorla's lecture is coincidentally titled *Reifying Representations*: 'coincidentally', as the choice for this present paper's title has been made independent, and long before Rescorla's lecture, which was held September 12th 2017 at UCL.
[43] See, for instance Barat, 2003: 812.

Importantly, and *pace* Rescorla, thinking of something as some 'thing' involves much more than countability. Typically, 'things' are thought to be characterized, not only by countability, but also by independent spatiotemporal existence, by being in principle perceivable, by having physical properties, by physical property-based identifiability, by localizability, by finiteness in space and time, by causal efficaciousness and so on. So reification with regard to non-things (actions and events, relations, norms, rules, qualities, contentful assertions etc.) refers to the practice of conceiving of these non-things as attributable with this list of properties. This is, at least, how I will understand reification.[44]

2.5.1 Reification through hypostatization

It is an undeniable fact of human nature that natural language profoundly influences the way we experience the world. The exact reach of this influence is of course a matter of debate, but that there *is* such an effect from language to cognition has not only been scientifically established, it is something we can easily accept merely by considering the fact that when a human being has reached a sufficient level of linguistic competence, she can no longer hear the utterances of its fellow language users in the same manner as before, nor can she see, after she has learned to read, certain shapes or combinations of shapes in the way she did at an earlier time; she has learned to see them, first of all, as letters, words or sentences, and it has become difficult, if not at all impossible to see these objects in the same way as before she acquired the reading skill. In short, gaining language means losing the pre-linguistic perspective of the animal or infant. The same physical objects (sound waves, shapes on a sheet of paper, hand gestures in sign language, braille, music notes, and so on) are being completely differently perceived before or after language, which seems proof enough of the

[44] In this sense, my use of the term reification is also different from Ludger van Dijk's, who uses the notion to denote the bigger psychological phenomenon of backwards causation. It would seem, however, that what van Dijk calls 'concretization' (which is one of three aspects of backwards causation) is precisely what I mean by reification. Terminological matters aside, van Dijk's analysis of the three-step phenomenon he calls reification deserves close attention as it lays bare uncritical tendencies in our explanations about behavior which require attention themselves. See van Dijk (2016).

fact that language has a profound effect on cognition, at least on our perceptual experience of the world. [45]

Conversely, however, the way we human beings experience the world has itself an influence on the nature of language. This becomes clear when we consider the grammatical categories of natural language. The fact that all known languages have nouns and verbs, for instance, is no arbitrary matter, but reflects something fundamental about human cognition, namely that we structure the world at some basic level in spatiotemporally extended objects (including other 'subjects'), and the activities these objects exhibit (including the actions of others). Perhaps not at its joints, but we were already carving up nature long before we acquired our language skills. Like other animals, we pre-linguistically structure the world a certain way, and our syntax reflects that fact. I take all of this to be fairly uncontroversial.

However, language also tends to lead a life of its own. All languages are rife with idioms, appearing on both the semantic and the syntactic level. For instance, sentences in which nouns or substantives literally refer to physical objects or 'substances' are much less current than one might suspect. So, a sentence like 'The cat is on the mat', in which both substantives refer to physical things (cat, mat), may be the handbook's preferred example of an assertion, it is hardly representative of the idiomatic complexities of ordinary language. Even the tritest of statements like 'I'm going to take a walk' becomes a deeply mysterious utterance to anyone insufficiently familiar with the English language. As a literal assertion, it seems to make little sense. For how does one literally *take a walk*?

As said, natural language practices abound with the nonliteral, but for my present purposes, I want to focus on a linguistic phenomenon of which the former example is already an illustration. The phenomenon I want to address is usually referred to as hypostatization. Hypostatization can be defined as the linguistic practice of placing non-substantive entities like actions or qualities in the

[45] It has been demonstrated that language guides thought (e.g. Ervin-Tripp 1967), has an influence on our concepts of time and space (e.g. Boroditsky 2001), and also affects memory (e.g. Loftus and Palmer 1974).

grammatical category of nouns, so as when an activity-indicating verb like 'to walk' becomes a noun in the expression 'taking a walk'. As such, hypostatization is an instance of the much broader linguistic phenomenon of grammatical derivation, which isn't necessarily a derivation to a noun; verbs or qualities can be derived *from* nouns as well (e.g. 'chairing' or 'friendly'), and we also find derivations from verbs to adjectives (e.g. 'talkative') or from adjectives to verbs (e.g. 'to sweeten')[46]. Hypostatization is a remarkably common form of derivation, and it can be found, at least in the English language, virtually everywhere. In fact, the expression 'it can be found everywhere in the English language' already contains two hypostatizations: first, the 'it' as referring to 'hypostatization', for the noun 'hypostatization' is itself an example of hypostatization; and second, 'the English language', for languages aren't things, and they certainly aren't things in which other things can literally be found. Crucially, however, what makes hypostatization such a unique form of grammatical derivation is that, unlike other derivations, the derived 'noun' term does not literally refer to a substance. In contrast, when verbs or adjectives are derived from nouns (e.g. 'hammering' or 'manly'), the terms *do* refer to an action or a quality respectively. Apparently, it is only with derived noun-terms that we encounter the phenomenon of an entity being referred to with a grammatical category to which it does not literally belong. In a sense, then, it is only here that grammar might set us on the wrong foot.[47]

Hypostatization is sometimes used as synonymous with reification. Here, however, I will consider both terms as referring to two different, though very closely related phenomena. I will take 'hypostatization' as referring to the just described linguistic phenomenon of categorical substitution, whereas I will use 'reification' as referring to the earlier explicated cognitive phenomenon of

[46] The phenomenon of linguistic derivation is, of course, not an exclusive feature of the English language. So, for instance, in Spanish we find, as an instance of a verb derived from an adjective 'verdear' (meaning something like 'becoming green'), or in Russian, as an instance of a verb derived from a noun, we find the verb 'solit', which means 'to cover with salt' ('sol' meaning 'salt'). These examples come from Lachlan Mackenzie (Mackenzie, private correspondence, see also: Mackenzie 2004).

[47] We may wonder why this is so. According to Mackenzie, this is probably related to the fact that only nouns can denote independently, whereas verbs or adjectives require a broader semantic-syntactic context. (Mackenzie, private correspondence).

conceptually treating non-things as things. Both phenomena are, as I've just said, closely connected in the sense that one can give rise to the other. Indeed, they form a good example of the abovementioned reciprocal influence between language and cognition. Here, I will be particularly interested in the 'language-to-cognition' direction, so in how hypostatization might give rise to reification.

It should be mentioned that in day-to-day conversations, the ubiquity of hypostatization rarely causes any difficulties, at least not between sufficiently fluent interlocutors. At worst, it might lead to a brief, often comical misunderstanding, which is usually quickly straightened out. So it might happen that we have to clarify we're referring to a movie, and not an actual person, when we tell a friend that we went to see 'The Man on the Train', or, citing Ryle's classic example, that the university isn't in fact one building. (Ryle 1949/2009: 6) What these examples show us is that, in addition to what the words of a sentence are supposed to mean, *the grammatical categories themselves are semantically relevant as well*. What we need to ask is this: why is it that the linguistic phenomenon of hypostatization can give rise to false beliefs, related to reification? This can only be so if the fact that something is represented by a noun is, regardless of its referential meaning, itself already a meaningful element. In other words, even if we don't understand what 'a walk' means, the fact that we recognize it as a noun is itself semantically relevant. It is precisely the fact that we already have some understanding of what it ordinarily means to be represented by a noun that explains the confusion in the first place. When something is represented in language by a noun, and we recognize it as such (because, for instance, it is preceded by an article), this brings with it certain ontological expectations, even if we have never heard the word before. If we come to believe that a noun is being used to denote a 'thing' or a substance or a material unity, these expectations include: independent spatiotemporal existence, having physical properties, physical property-based identifiability, quantifiability, localizability, finiteness in space and time, having causal efficaciousness, and all the other characteristics we usually associate with what it means to be 'a thing'. When these expectations are turned into actual beliefs about the referent of the

noun, we have reified the referent. And although hypostatization rarely generates practical or theoretical problems, sometimes it does, even for a native speaker. It is to the merit of philosophers like Ryle, and Husserl before him, to have identified these problems, either as category mistakes (Ryle) or a *metabasis eis allo genos* – change to another genus (Husserl)[48].

My point with regard to the notion of representation, then, is that the fact that mainstream cognitive science discourse has adopted, and continually reiterates the habit of referring to 'internal representations' in substance terms, has a psychological effect on the way these hypothetical entities are conceived of ontologically. It gives the reified notion a *prima facie* credibility which it does not deserve, based on our previous analysis. And the scientific mind is susceptible to the psychological effects of language as well, even though some scientific minded readers will doubtlessly find that the above does not apply to them.[49] They are invited to interpret the above as a cautionary reminder not to fall into the pitfalls of ordinary language. The take-home message here is that adopting a reified notion of representations comes with the responsibility of being able to give language independent reasons for this reified conception, and not simply assume it based on the already installed practice of grammatically referring to these entities by a substantive.

[48] As Reynaert (2015) points out, in his *4th Logical Investigation*, Husserl carefully distinguished between the nonsensical and the absurd. On his account, 'metabasis eis allo genos' or 'change to another genus' gives rise, not so much to nonsensical, but to absurd notions. Arguably, Husserl would consider the reified notion of representation as employed within contemporary mainstream cognitive science as technically absurd. See Reynaert (2015) and Thomasson (2013) for further discussion.

[49] An objection may be that representation is a functional notion, and that 'representation' first of all refers to the activity of, say, a neural state. It may be argued that it is because certain neural states are in the business of, or, to use a more common rhetoric, fulfill the role of representing, that we may correctly identify them as representations. Note, however, that this functional-role interpretation does not in any way escape the reified picture, and all the problems associated with it. It is still some physical object that is supposed to be fulfilling the role of representing and it still needs to be shown in a naturalistically credible way how such an object (presumably some internal neural structure) can literally be fulfilling the role of representing or standing for something else.

2.5.2 Reification through ambiguity (1): 'representation' as an ambiguous term

The second reifying contingency is related to the fact that, at least in some languages, the term 'representation' does indeed literally and unproblematically pick out ordinary physical objects. In English, the term not only refers to contentful entities like words, sentences, gestures, or other things to which we attribute semantic content; here, the noun 'representation' is also understood as synonymous to 'picture', 'drawing', 'image', 'illustration', 'sculpture' and other objects which can be unproblematically referred to as things. In these cases, the use of the word representation seems to be motivated, not solely by the idea that these objects 'stand for something', or 'mean something' , or 'are about' states of the world, but more simply by the idea that they quite literally re-present – in the sense of 'making present again' – something perceptually (mostly visually). These objects are categorized in relation to our cognitive ability to recognize something – which is typically absent – in something else, based on some sufficient degree of perceived similarity. This is something not only humans can do, but other animals as well. We are biologically wired to 'structure' (though, of course, not in the sense of classification) the world based on perceivable structural similarities. We are, however, *not* biologically wired to actively manipulate and utilize this natural ability by generating artefacts which have structural similarities to other things ourselves. A dog may start barking at a life-sized sculpture of a dog, but only humans will actually craft such a thing. In English, we call these artefacts 'representations', and as said, in these cases, the noun does indeed quite unproblematically pick out physical things. But note that, when it comes to this type of representation, there is no need to invoke truth evaluable content whatsoever. Remember that, for something to qualify as a contentful representation, according to Rescorla, it must be "semantically evaluable with respect to such properties as truth, accuracy, and fulfillment". To the extent that the noun 'representation' refers to physical objects simpliciter, like pictures or sculptures, Rescorla's criterion doesn't apply. Thinking that it *does* is misguided, as I will try to show below.

Within cognitive science, we find alternative notions of representation that are modeled on precisely the kinds of ordinary representations discussed above. Rather than thinking of representations as modeled on linguistic entities, some authors have adopted a view of representations that sees these entities as more akin to pictures, maps or models, focusing on the structural resemblance between representations and their targets (e.g. Gladziejewski 2015, O'Brien & Opie 2015, Ramsey 2007, Waskan 2006, Braddon-Mitchell & Jackson 1997, McGinn 1989, Craik 1967). Despite their differences, what is retained in all of these conceptions is the idea that, in order for something to count as a representation, there have to be some kind of correctness conditions in play. The problem with these views is that the ordinary external representational objects on which these theoretical notions are modelled do not seem to require the involvement of any semantically evaluable content whatsoever, and in any case, we don't need the idea of 'semantic evaluability' to properly characterize these objects. Although I find many elements in McGinn's theory of mental content problematic, he is surely right in stating that "[s]entences have semantic properties …; models do not – any more than maps or tree rings do." (McGinn 1989: 181) It is true that objects like pictures, maps and models can be said to be more or less accurate, but that doesn't necessarily mean they are accurate or inaccurate in any content involving sense. Take a picture, for example. A picture, *qua physical object*, has no truth, accuracy, correctness or any other conditions of satisfaction, because, considered merely in its physical nature, no object does. From an objective perspective, a picture just is a physical structure of lines and shapes, and perhaps colors, in which we (and not necessarily everyone) happen to recognize something else. Of course, nature is full of physical structures in which we might recognize something else (think of the ubiquity of pareidolic objects), yet these don't qualify as pictures. In contrast to actual pictures, with these natural structures, the resemblance is merely coincidental, whereas with pictures, it is intentional. Pictures aren't just physical structures, like baseballs, they are normatively constituted artifacts that are *supposed* to bare a resemblance to whatever is being depicted. In this sense, they have certain normative conditions. A picture can indeed be more or less accurate, depending

on the picture's goal, for example creating a physical structure that sufficiently resembles the depicted. If that is what the picture is supposed to be doing, then, quite trivially, the picture has certain conditions that can be spelled out in terms of similarities, but only because *we subject the object* to certain norms. Furthermore, to the extent that the picture is supposed to be aesthetically pleasing, it has aesthetic conditions as well. But why would talk of an object having certain conditions of satisfaction require the involvement of semantic content? As far as I can see, this would only be the case if the relevant conditions would be truth conditions, but these are precisely the kind of conditions we don't need when considering pictures, maps, models or any other physical object we refer to as a representation (with the exception, of course, of spoken or written assertoric sentences). Why, then, do we keep seeing these objects reappearing in discussions of assumed content-bearing mental representations? In recent years, many theorists have turned away from the idea of modelling mental representations on linguistic structures. Yet, at the same time, these authors want to hold on to the idea of semantic content. But this is precisely what you can't get if you conceive of mental representations in terms of structural similarity (isomorphism or homomorphism). Reverting to accuracy or correctness conditions seems to be a halfway house which, on closer inspection, causes more problems than it solves: thinking of internal representations as pictures, models or maps, rather than assertoric sentences leaves us, from a naturalistic perspective, with the worst of both worlds. On the one hand, like sentences, these objects are just as much socio-normatively constituted, so the problem of reducing the socio-normative to the objective non-normative remains; on the other hand, contrary to assertoric sentences, there is no reason to ascribe truth evaluable content to them, so they are of no use as a model for mental representations understood as internal entities that somehow say how things stand with the world such that they could be different. Neither pictures, models, maps, nor any other isomorphic structure does this. Our assertoric judgments about these structures do. For instance, a sufficiently accurate model of the solar system, because of a sufficient structural similarity, helps us in making true assertions about, say, the relative position of the planets. These model based

assertions are, of course, truth evaluable, contentful semantic entities. But, just as with the example of the fire alarm, that doesn't mean that the model itself is carrying truth evaluable content as well. The idea that it does seems to be again an instance of what I referred to above as metonymy or displacement.

To sum up: the second reason why it might seem justified to treat 'internal representation' as referring to an actual thing comes from the fact that we do indeed sometimes use 'representation' to appropriately refer to actual things (pictures, models, maps…). But to the extent that this notion picks out physical structures that bear structural similarity to something else, it can be entirely cashed out in non-semantically laden terms. Like any other physical structure, these things are not, in themselves, truth-evaluable. Recall Hutto & Myin's 'Covariance Does Not Constitute Content principle'. "And neither does structural similarity", we might add.

2.5.3 Reification through ambiguity (2): 'state' as an ambiguous term

There is another way in which ambiguity can come to serve a facilitating role when it comes to the reification of mental representation. Next to the intrinsic ambiguity of the term 'representation' itself, within mainstream cognitive science, there is another closely associated notion at work which is equally ambiguous, but which endorses reification in a perhaps more subtle way. The notion I'm thinking of here is that of a state. Cognitive science literature is rife with talk of mental states, brain/neural states, intentional states, representational states and so on. What is typically overlooked here, however, is that the meaning of 'state' changes considerably depending on the combinations it is used in. In this regard, it is worth to bring up the oft-quoted § 308 of Wittgenstein's *Philosophical Investigations*:

> How does the philosophical problem about mental processes and states and about behaviourism arise? – The first step is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometime perhaps we shall know more about them – we think. But that is just

> what commits us to a particular way of looking at the matter.... (The decisive movement in the conjuring trick has been made, and it was the very one that we thought quite innocent.)

The quote's particular relevancy for our discussion should be clear, as Wittgenstein is explicitly referring to states here. More than half a century later, cognitive science has still left the nature of states 'undecided', allowing theorists to use 'mental states' and 'brain states' interchangeably. But what, on closer inspection, allows us to treat these various uses of 'state' as equivalent, or even similar? After all, the way 'states' are ascribed to brains is entirely different from the way 'states' are ascribed to a person. When we are ordinarily speaking of a mental state, we are using the term 'state' in the meaning of 'condition'. 'Mental states' first of all pick out the different psychological conditions in which a person (or an animal) can be said to be. When we ordinarily inform about mental states, we first of all want to know how *someone* is doing. What we are *not* interested in is a certain physical *configuration*, which is a whole different sense of the term 'state'. In a second, perhaps less ordinary sense, a person's mental state might also be used to refer to, not *how* the person is doing, but *what* he is doing in his mind. For instance, when we ascribe assertoric thoughts to a person, we could say that the person is in a certain mental state, namely that of thinking. And to the extent that thinking is a representational activity with intentionality, we might say that he is in an intentional mental state. But notice how, also in common speech, the *is* has changed into *has*. Rather than saying that a person *is* in an intentional state, we tend to say that he *has* intentional states. Saying that a person is thinking or saying that he has thoughts is in ordinary speech perfectly interchangeable. However, the move from *being* in an intentional state to *having* an intentional state is a step towards reification, and one that is unmotivated by science. The ambiguity of 'state' further completes the reifying process. With some conceptual leniency 'brain states' can indeed be said to be things, or at least configurations of things (neurons). But now, the conjuring trick is easily accomplished. Because brain states are physically localizable things with physical properties, mental states in general, and intentional states in particular, must also

be like that, hereby conveniently overlooking the ambiguity of 'state'. Furthermore, since they are now conceived of as 'thing-like', they can also be *had*. *Being* in a representational state (thinking) has now become synonymous with *having* a representation (a thought). And, so it is assumed, having a representation *just is* having a brain state.

2.5.4 Reification through the notion of properties

As we've seen, cognitive science's generic notion of internal representations as truth evaluable entities (entities saying how things stand with the world such that they could be different) is modeled on our pre-theoretical familiarity with such entities. But this means that internal representations are modeled on assertions, claims, judgments, and other linguistic items with truth evaluable content, for these are the *only* entities that can literally be said to have truth conditions. The question then becomes: does it make sense to think there can be assertions in the brain? Now, perhaps one may object to this formulation and say that, literally speaking, it would be absurd to expect to find assertions in the brain. He or she might add that perhaps we might find something sufficiently analogous to assertions, so that we can still say that it shares the relevant property of semantic evaluability with actual assertions, without it literally being an assertion. This line of reasoning, however, hinges on the assumption that truth evaluability is, first, a kind of *property*, and, second, a property that can be ascribed, not only to certain linguistic entities like assertions, but to extra-linguistic entities as well (neural states, for instance). As I want to show, invoking property talk in relation to semantic issues tends to promote reification.

Within analytic philosophy, the use of the notion of property abounds. Likewise, talk of semantic or representational *properties* has become an almost routinous affair. Yet, as anyone with a background in analytic philosophy knows, the philosophical notion of property is highly debated. In its most ordinary, more narrow sense, 'property' is used to refer to an object's physical features. Roughly, they are the 'things' we would be mentioning if we were to describe the object.

However, in its wider sense, 'property' is understood as synonymous with truthful predication. Anything that can be truthfully predicated of an entity can be said to be a property of that entity. So when it is asserted that an assertion like 'the cat is on the mat' has semantic properties, what is usually meant is that its semantic evaluability – i.e. the entity's having truth conditions – can be truthfully predicated. Of course, no one believes that the sentence 'the cat is on the mat' has the property of being truth evaluable merely in virtue of that sentence's physical properties. But note, however, that regardless of whether one adopts the narrow notion of property or the wide notion preferred by analytic philosophy, what remains in place is the idea that 'things' like assertions can be incorporated in the same metaphysical picture we use to classify the whole of nature, namely in terms of objects and their properties. On this construal, semantic properties may be different from ordinary, material properties, but the ontological foundation on which it is construed remains untouched, namely the object/property distinction itself, and its logical reflection in the subject/predicate distinction. By simply invoking the notion of property in relation to semantic issues, one is framing these issues in a metaphysical picture in which they might not belong. As already discussed, using nouns to refer to abstract entities like assertions is one thing (hypostatization), but to conceive of these 'things' as *being* physical objects with properties is another (reification).[50] Yet it is precisely this reified conception of representations as physical objects with properties we keep encountering in cognitive science literature, not only in the more traditional computationalist approaches (e.g. Newell and Simon's *physical symbols* system), but in newer approaches as well. Take the signal system approach to the brain, for example. Here, the semantic notions on duty are codes, information and "signals running around a very complicated signaling network" (Skyrms: 2010). But no matter what nouns are being used to denote the

[50] Following Wittgenstein, McGinn (1989) makes a very similar point when he warns us against "…assimilating intentional properties to the properties characteristic of substances." (McGinn 1989: 29 ft. 40). And further he reminds us: "Wittgenstein, of course, had the idea that there is a persistent and rooted tendency to model the mind on the world of material objects. It is not as if we come to see that this is false and there's an end to the matter. Prolonged therapy may therefore be needed to dislodge wrong philosophical conceptions. We might have an internal fight on our hands." (McGinn 1989: 30 ft. 41)

supposedly semantic-content-carrying entities, the same familiar picture keeps recurring, namely that of an *object* with both physical and semantic *properties*. Consider how Rosa Cao, for instance, discusses the role of the receiver in a signal system:

> To interact with signals in the right way (so as to be a receiver of semantic information) is, roughly, for the receiver to have some degree of flexibility in its response to a signal. At first pass, this means that receivers will be aptly described as acting (at least in part, but perhaps primarily) on the basis of the semantic properties of a signal*, in addition to its material ones*. (Cao 2012: 53, my italics)

Conceiving of semantic properties as being on a par with material properties is one more expression of reification. It results from the mistaken assumption that truth evaluable representations can be conceived of as physical objects with intrinsic properties, both material and semantic ones. But it is a mistake to think that semantic properties are somehow in line with material properties. In his paper *Naturalizing Representational Content*, Nick Shea provides us already in the opening lines with a clear instance of this error. He writes:

> Some things in the world have semantic properties. Spoken and written sentences are paradigm cases. They are perfectly ordinary particulars in the causal order: ink marks on the page and vibrations in the air. But they also have more exotic properties: they can be true or false, or, in the case of imperatives, they can be satisfied, or go unsatisfied. (Shea 2013: 496)

Spoken and written sentences are *not* perfectly ordinary particulars in the causal order. They are, as we've argued, socio-normatively constituted. However, all of this does not mean that talk of semantic properties is always and everywhere problematic. When an analytic philosopher like Davidson, for instance, uses the vocabulary of semantic properties, he does this exclusively in relation to abstract linguistic entities like sentences. As we've already seen, the wide sense of property (true predication) allows for this: since it can be truthfully predicated or asserted of an assertoric sentence that it has truth conditions, it is correct to say that the sentence has this property. But one can't simply transfer this language-philosophical approach to physical objects, whether they are 'ink marks on the

page', 'vibrations in the air', hand gestures or neural structures. Again, semantic properties are not on a par with physical properties. Conflating these very different properties is one more factor which facilitates and reinforces a reified understanding of internal representation.

2.5.5 Reification through the vehicle/content metaphor

One of the most reiterated, yet at the same time most underexamined idioms in cognitive science literature is the commonplace distinction between a vehicle and its content. As already mentioned, on a standard interpretation, internal representations are conceived of as content carrying vehicles and most theorists simply assume the distinction without explicitly thematizing it (for an exception to this, see for instance Hurley 1998). There is discussion about when exactly the distinction got introduced into philosophy of cognitive science (Dennett 1969 appears to be one of the earliest sources[51]), but it is important to note that the conceptual distinction itself already existed long before people like Dennett and Millikan[52] started to use it in discussions about consciousness or the naturalization of semantics. The idea of thinking about content as being carried by physical vehicles is not born within philosophy of cognitive science, and it certainly isn't motivated by empirical research (despite what some theorists seem to belief[53]). We have good reasons to assume that the distinction long predates cognitive science, and that it has emerged as a result of certain specific historical developments, namely the practice of writing, printing and distributing books. To my knowledge, the rather obvious link between the vehicle/content distinction and these historical events has never been made explicit within cognitive science

[51] In *Content and Consciousness*, Dennett writes: "The crucial point that emerges from this is that the candidates for *vehicles of content* or significance in the brain are compound." (Dennett 1969: 56; my italics)

[52] See, for instance, Millikan's *Content and vehicle* (1993)

[53] See for instance Manzotti & Pepperell 2013. The authors criticize an anonymous referee, referred to as *A*, who apparently takes the content/vehicle distinction to be empirically established: "For *A*, the nature of mental vehicles, and thus their separation from mental content, is an empirical matter rather than a terminological one. Yet (…) we wonder where the empirical evidence is for this distinction existing anywhere other than the minds of those who believe in it?" (Manzotti & Pepperell 2013: 368)

literature. Doing so, however, quickly reveals one more contingency supporting the reified picture of representation.

Before anything else, it should be again emphasized that the vehicle/content distinction is first of all a *metaphor*. Or, rather, it is a *double* metaphor in that both the term 'vehicle', as well as 'content' are used in a non-literal sense. For our discussion of reification, however, it is especially the metaphorical notion of content that is of interest here, and not so much that of a vehicle (vehicles are about as thing-like as it gets). In its literal sense, 'content' is that which is being contained by a *volume* of some kind. It is, then, not difficult to see how the literal idea of content came to be associated with, and ultimately metaphorical for, meaning, considered in relation to writing, and the practice of writing books in particular. In fact, the earliest reference we find to content in relation to meaning connects the word to books. In the Oxford English Dictionary, the first mention of content in this sense dates back to 1481:

> Here endeth the table of the content and chapytres nombred of this present book.[54]

It is also no coincidence, then, that we refer to books as *volumes*. And just as amphora's, cups, boxes and other volumes can contain fluids or other materials, so too can books contain 'material', i.e. meaning. Furthermore, if we want to know what the book is about, we typically do so by asking what's *in* it. This image is not restricted to the English language. We find that this metaphor of meaning as content not only has roots in Latin (*continere)*, but in other Indo-European languages as well (German: 'Inhalt', Dutch: 'Inhoud', Danish: 'indholdet'...). And of course, since books – like other volumes – are the kind of things that can be carried around, the idea of it serving as a vehicle for its content follows quite naturally. The problem, however, is not that this picture reifies the 'carriers' of meaning, but meaning itself. It is, after all, hard to see how something can be *in* a volume and be carried around by a vehicle without it being a physical thing itself. In other words, the idea of thinking about meaning as something that can be

[54] Caxton tr. *Siege & Conqueste Jerusalem*, 1481, edition of 1893.

contained by a volume and transported by a vehicle supports a reified understanding of meaning. This does not have to be problematic in itself, as long as we keep in mind that we are dealing with a metaphor here, and that the question as to how the metaphor could be cashed out in naturalistically respectable terms still stands. Nevertheless, within contemporary cognitive science, and especially within standard informational-computational approaches, the idea that the brain performs computational operations on content carrying vehicles (representations) is, as a rule, taken quite literally.

2.6 Reification and scientific explanation

The focus of the discussion so far has been on conceptual elements which plausibly sustain and reinforce the reified concept of representation, yet which are themselves not grounded in scientific practice. I have therefore referred to these elements as 'contingencies'. It should be noted, however, that one important contribution to the reification of representation comes from *within* scientific practice, and is therefore not properly labeled a contingency. As already indicated at the beginning of this paper, one major reason for supporting the reified picture of representation is that it renders the notion deployable within the larger framework of causal explanation. It is our understanding of what it means to be an objective scientific explanation that already itself motivates the idea of internal representations as object-like entities with, therefore, object-like causal properties. Viewed in light of the demands of a scientific explanation, it isn't hard to see what makes the reified construal an attractive hypothetical posit. It is an attempt of tying in two threads which are deemed indispensable for the explanation of psychological phenomena. On the one hand, cognitive scientists want to hold on to the idea that the assumed contents of our mental states are explanatorily essential. On the other hand, it has become a central assumption of scientific theorizing that, for something to be a properly objective explanation, the explanation must be – in the final analysis – causal. The reified notion of internal representation, which, as we've seen, is inherently dual in nature,

accommodates these demands by stipulatively uniting and incorporating both elements. On closer examination, however, we come to see how incoherent such an idea really is. When we think about what it means for an account to qualify as an objective causal explanation, we see that one of the conditions is that we precisely *abstract away from*, or *leave out content*. Causal explanations are objective in the sense that they explain how and why things are the way they are regardless of how subjects think or feel about them. From the objective scientific perspective, content simply *cannot* play an explanatory role, but is itself in need of a causal explanation. Conversely, explaining why someone acted the way she did in terms of content (e.g. "Because she believed it was the right thing to do.") requires us precisely to *not* give an account in terms of cause and effect. Both explanations do not have to be incompatible, but you can't have them *both as one explanation*, even if it should turn out that one explanation (presumably the content-invoking one) is reducible to the other (presumably the causal one). The reified notion of representation, however, seems to think it can do just that by stipulatively fusing them into one hypothetical entity: an objective symbol with causal power. In a sense, then, internal representations are the reified and intracranial version of what Sellars once described as the fusion of two different perspectives into one "stereoscopic view" (Sellars 1963: 5). I'll explain this – admittedly suggestive – claim in a bit more detail.

In his *Philosophy and the Scientific Image of Man*, Sellars famously argues that the contemporary philosopher is confronted with two very different, yet "equally public, equally non-arbitrary conceptions of man-in-the-world" (Sellars 1963: 5). As is well-known, he refers to these two conceptions as the *Manifest Image* and the *Scientific Image* respectively, and he takes it to be one of the great challenges of philosophy to understand how both images hang together. For my purposes, I'll define the Manifest Image as that perspective from which the world appears as what it is *for us*, or more precisely, *for a community of persons*. It is our everyday-perspective in which a collection of H_2O molecules first appears as drink water, or black ink lines on a piece of paper as saying that someone's going to be back in 5 minutes. The Scientific Image can be defined in contrast to the Manifest Image

in that, from the scientific perspective, the world does not appear as what it is for us, but how it is in itself, that is, how it is *objectively*. Within the Scientific Image, drinkable water appears first of all as a collection of H_2O molecules, or a note saying that someone will be back in 5 minutes as black ink lines on a piece of paper. I'm oversimplifying matters here a bit, but this doesn't affect the point I'm trying to make. For it seems rather clear that those things we refer to as representations, that is, objects which are said to be carrying content or meaning, are firmly confined within the Manifest Image, and have – quite literally – no meaning within the Scientific Image. Objects we call representations (linguistic items, traffic signs, maps, models…) but also all other normatively constituted things like coins, chess pieces or, indeed, baseballs, simply lose their meaning, and therefore their identity, when viewed through the objective, disengaged lens of the scientist interested in providing causal explanations. To be more precise, an essential property of *any* representation (whether it is public or intracranial) is that it has content. But content is precisely what we lose when we look at these entities from the perspective of the Scientific Image. In short, then, the idea of a content carrying representation makes no sense outside the Manifest Image. It is perhaps not impossible to *explain* the existence of contentful representations in objective-causal terms, but it does seem conceptually confused to *use* contentful representations within such objective-causal explanations. Yet this is precisely what the notion of internal representation is supposed to be doing in causal-mechanistic explanations of mind and cognition, for instance in computationalist explanations. In other words, what is seen by philosophers like Sellars[55] as perhaps the greatest philosophical challenge (fusing the Manifest and the Scientific Image together in one stereoscopic view), is by mainstream cognitive science already considered a *fait accompli* by postulating internal content carrying entities which are at once causally explanatory. The reified notion of representation wants to provide a 'best of both worlds' explanation by combining essential elements from both explanations within the Manifest Image, as well as

[55] I could just as well have referred to Peter Strawson, who also highlights the distinction between two possible, yet different perspectives on the world. What Sellars captures in terms of Manifest vs. Scientific Image is very similar to what is by Strawson referred to as the possible "occupying" of two "alternative standpoints". See Strawson 1985: 55.

the Scientific Image, i.e. content and physical causation, respectively. This requires the postulation of a physical entity (to accommodate the causal efficaciousness condition) that somehow carries content in a non socio-normatively constituted sense. But this fusion can only be a confusion. In light of the above, it should be clear that the idea of such *things* (reification) existing in the head or elsewhere appears to be incoherent. Ultimately, the incoherent nature of the idea is the result of confusing, or rather, conflating a prescriptive socio-normatively constituted entity – within the Manifest Image – with an objective-descriptive one – within the Scientific Image. Indeed, this is the same as saying that the reified notion of internal representation is the result of a category mistake.

2.7 Concluding remarks

Speaking of abstract normative entities like representations in terms of thing-like particulars with thing-like properties does not have to be always and everywhere problematic. However, when reification leads to an actual change in our beliefs about these entities' ontological status, problems might arise. This is especially the case when the reified entity becomes an indispensable part of a supposedly scientific explanation. For the kind of explanations we find in representational-computational theories of cognition, it is essential that representations are thing-like entities in that they are supposed to have both causal properties, as well as the property of being spatiotemporally localizable (they are, after all, *internal* representations). At the same time, these thing-like entities are also supposed to have the property of carrying semantically evaluable content, where the latter is assumed to have causal relevance. As the above de-reification has shown, outside of a socio-normative context, such entities cannot be found. We *can* make sense of thing-like entities carrying semantically evaluable content, but only in connection to our socio-normative practice of giving and asking for reasons. However, nothing *intrinsically* has a certain content. In an important sense, then, the descriptive fact that some thing can be said to have a certain content follows

from a social prescriptive fact. Nevertheless, the idea that a physical entity can be said to carry a specific content in and of itself, in virtue of its properties, is still a hard-wired assumption within cognitive science theory. In addition, I've tried to show that presupposing a reified notion of representations is not so much grounded in scientific research as it is determined by contingencies in our ways of speaking and thinking. In the final analysis, mainstream cognitive science's central explanatory posit turns out to be incoherent. So as long as cognitive science's engine keeps running on internal representations, it might perhaps maintain its speed, yet it would keep on heading in the wrong direction.

References

Bakhurst, D. (1991). *Consciousness and Revolution in Soviet Philosophy*. Cambridge: Cambridge University Press.

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs 28*: 801–831. Gender and Science: New Issues. The University of Chicago Press.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues 13*, 31–45.

Boghossian, P. (2005). Is meaning normative? In *Philosophy–science–scientific philosophy*, ed. C. Nimtz and A. Beckermann, 205–218. Paderborn: mentis.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology 43*: 1–22.

Braddon-Mitchell, D., and Jackson, F. (1996). *Philosophy of Mind and Cognition.* Oxford: Blackwell.

Brandom, R.B. (1994). *Making It Explicit Reasoning, Representing, and Discursive Commitment*. Cambridge MA: Harvard University Press.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology & Philosophy 27*: 49–71.

Cash, M. (2008). Thoughts and oughts. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action 11*: 93-119.

Cash, M. (2009). Normativity is the mother of intention: Wittgenstein, normative practices and neurological representations. *New Ideas in Psychology 27*: 133–147.

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

Craik, K.J.W. (1943/1967). *The Nature of Explanation*. Cambridge University Press.

Davidson, D. (2005). *Truth and Predication*. Cambridge, Mass.: Belknap Press.

Dennett, D. C. (1969). *Content and Consciousness*. Routledge and Kegan Paul plc.

Egan, F. (2012). Representationalism. In Margolis, E., Samuels, R. & Stich, S. (eds.), *The Oxford Handbook of Philosophy and Cognitive Science*. Oxford University Press.

Ervin-Tripp, S. (1967). An Issei learns English. *Journal of Social Issues 2*: 78–90.

Fodor, J. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive science. *The Behavioral and Brain Sciences 3*: 63–73.

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

Fodor, J. (1987). Meaning and the world order. In *Psychosemantics*, 97–133. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, Z. (1981). How direct is visual perception? Some reflections on Gibson's 'ecological approach'. *Cognition 9*: 139–196.

Gallagher, S., & Miyahara, K. (2012). Neo-pragmatism and enactive intentionality. In: Schulkin, J. (ed.) *Action, Perception and the Brain. New Directions in Philosophy and Cognitive Science*. Palgrave Macmillan: London

Gibbard, A. (2003). Thoughts and norms. *Philosophical Issues 13*: 83–98.

Gibbard, A. (2005). Truth and correct belief. *Philosophical Issues 15*: 338–350.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric 40*: 63–90.

Godfrey-Smith, P. (2014). Signs and symbolic behavior. *Biological Theory 9*: 78–88.

Hattiangadi, A. (2006). Is meaning normative? *Mind & Language 21*: 220–40.

Haugeland, J. (1998). *Having thought: Essays in the metaphysics of mind*, 305–361. Cambridge, MA: Harvard University Press.

Hume, D. (2002/1739-40). *A Treatise of Human Nature*. 2nd ed. Norton, D.F., and Norton, M.J. (eds.) Oxford: Oxford University Press.

Hurley, S. (1998). Vehicles, contents, conceptual structure, and externalism. *Analysis 58*: 1–6.

Hutto, D. D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Hutto, D.D., and Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Hutto, D. D., and Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

Hutto, D. D., and Satne, G. (2015). The natural origins of content. *Philosophia 43*: 521–536.

Kalish, C. (2005). Becoming status conscious: Children's appreciation of social reality. *Philosophical Explorations 8*: 245–263.

Lewis, D. K. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Loftus, E. F. and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior 13*: 585–589.

Luntley, M. (1999). *Contemporary Philosophy of Thought: Truth, World, Content.* Oxford: Blackwell Publishers Ltd.

Mackenzie, J.L. (2004). Semantic categories and operations in morphology I: entity concepts. In: Geert Booij et al. (eds.), *Morphology. An International Handbook on Inflection and Word-Formation*. Vol 2. Berlin: de Gruyter. 973–983.

Manzotti, R., and Pepperell, R. (2013). Denying the content–vehicle distinction: a response to 'The New Mind Revisited'. *AI and Society 28*: 467–470.

McGinn, C. (1989). *Mental Content.* Oxford: Basil Blackwell.

Miłkowski, M. (2015). The hard problem of content: Solved (long ago). *Studies in Logic, Grammar and Rhetoric 41*: 73–88.

Millikan, R. G. (1993). Content and vehicle. In N. Eilan, R. McCarthy, B. Brewer (eds.). *Spatial Representation*. Oxford: Blackwell. 256–268.

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery 19*: 113-126.

O'Brien, G., and J. Opie. 2015. Intentionality lite or analog content? *Philosophia 43*: 723–730.

Orlandi, N. (2014). *The Innocent Eye: Why Vision is not a Cognitive Process*. Oxford University Press.

Pylyshyn, Z. (1980). Computation and cognition: issues in the foundation of cognitive science. *The Behavioral and Brain Sciences 3*: 111–132.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Rescorla, M. (2016). Bayesian Sensorimotor Psychology. *Mind & Language 31*: 3–36.

Reynaert, P. (2015) Does naturalism commit a category mistake? *Bulletin d'Analyse Phénoménologique 9*: 1–20.

Rosenberg, A. (2011). *The Atheïst's Guide to Reality: Enjoying Life without Illusions.* W.W. Norton & Company, Inc.: New York, London.

Rowlands, M. (2006). *Body Language: Representation in Action*. MIT Press.

Rowlands, M. (2014). Arguing about representation. *Synthese.* Doi:10.1007/s11229-014-0646-4

Ryle, G. (1949/2009). *The Concept of Mind*. London: Hutchinson.

Sellars, W. (1962). Truth and "correspondence". *The Journal of Philosophy*, *59*: 29–56.

Sellars, W. (1963). Philosophy and the scientific image of man. In *Empiricism and the Philosophy of Mind*, 1-40. London: Routledge & Kegan Paul Ltd.

Shapiro, L. (2014). Radicalizing Enactivism: Basic Minds without Content, by Daniel D. Hutto and Erik Myin. (Review). *Mind 123*(489), 213–220.

Shea, N. (2013). Naturalising representational content. *Philosophical Compass 8*: 496–509.

Skyrms, B. (2010). *Signals: Evolution, Learning and Information.* Oxford: Oxford University Press.

Strawson, P.F. (1985). *Skepticism and Naturalism: Some Varieties*. London: Methuen & Co. Ltd.

Thomasson, A. L. (2013). Categories, *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/categories/>.

Tonneau, F. J. (2011/2012). Metaphor and truth : a review of representation reconsidered by W. M. Ramsey. *Behavior and Philosophy (Online) 39/40,* 331-343.

van Dijk, L. (2016). Laying down a path in talking. *Philosophical Psychology 29*: 993-1003.

Waskan, J. (2006). *Models and Cognition*. Cambridge MA: MIT Press.

Wittgenstein, L. (2009/1953). *Philosophische Untersuchungen/ Philosophical Investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte, Revised fourth edition by P. M. S. Hacker and Joachim Schulte. Wiley-Blackwell: West-Sussex.

3 Off-line cognition as representation-hungry? Turning representation-hungry problems on their heads

Abstract

Nowadays, many working in the field accept that we do not need to invoke internal representations for the explanation of on-line forms of cognition, but when it comes to explaining higher, off-line forms of cognition, it is widely believed that we must fall back on internal-representation-invoking theories. In this paper, I want to argue that, contrary to popular belief, there really is no compelling reason for assuming that non-representationalist theories are, as a matter of necessity, limited in scope. I will show that Clark and Toribio's influential argument in terms of 'representation-hungry' vs. 'non-representation-hungry' cognition can hardly be considered an argument at all. On closer inspection, we'll see that the claim from representation-hunger is built on a conflation of the level of description with the level of explanation. As we'll also see, this conflation is fairly common, both within, as well as outside the borders of cognitive science theory: overlooking the crucial distinction between representation as a descriptive notion and representation as an explanatory posit has, for instance, also caused authors to mistakenly identify Husserl as the father of contemporary representationalist cognitive science. I will argue that, to the contrary, Husserl's phenomenological investigations are not at all recuperable within today's framework of mainstream cognitive science. Even more, drawing on phenomenology, it can be argued that off-line cognition isn't so much representation-hungry, but that rather the opposite is the case, i.e., that representation is itself a cognitive phenomenon that needs to be understood as depending on certain forms of off-line cognition being already in place.

3

Off-line cognition as representation-hungry? Turning representation-hungry problems on their heads

3.1 Introduction: brains and/as computers

Within cognitive science, not many ideas are met with such an absolute divergence of opinion than the claim that brains are computers. There are those who accept the idea as both a literal and theoretically fundamental truth. For these theorists, cognitive science *just is* the study of the brain as a kind of computer. Therefore, approaching cognition means approaching it computationally, for cognition *just is* the computational processing of information.

Others tend to take a more cautious approach. They see the brain-as-computer idea rather as a useful metaphor. Instead of thinking of brains as literally *being* computers, these theorists prefer the idea that brains are *like* computers. Whether the analogy will ever turn out to be something more remains to be seen, but for now, these theorists find the idea acceptable as a helpful heuristic tool.

Then there are others still, who reject the brain-as-computer claim entirely. According to them, brains are not computers, nor are they *like* computers. The idea is deemed completely misguided and standing in the way of real scientific progress. Research psychologist Robert Epstein, a staunch opponent of the view that brains are computers, has recently emphasizes in a much discussed article[56] that the practice of modeling brains on computers is best understood as an historical contingency. Like others before him[57], Epstein argues that, from a

[56] See https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

[57] Rodney Brooks and Charles Randy Gallistel, for instance, have on several occasions also emphasized the remarkable link between our metaphors for the workings of the brain and contemporary technological developments.

historical point of view, the tendency to try to understand the human mind in terms of our latest technological achievements is as old as our theorizing about the mind itself. Epstein underpins this claim with recent work by George Zarkadakis. In his *In our own Image* (Zarkadakis 2016), the latter discusses the various metaphors humans have historically been employing to get a better grip on the nature of mind and cognition. Strikingly, these metaphors appear to be all drawn from relatively contemporaneous technological developments. I'm quoting Epstein's discussion of Zarkadakis at length:

> The invention of hydraulic engineering in the 3rd century BCE led to the popularity of a hydraulic model of human intelligence, the idea that the flow of different fluids in the body – the 'humours' – accounted for both our physical and mental functioning.…By the 1500s, automata powered by springs and gears had been devised, eventually inspiring leading thinkers such as René Descartes to assert that humans are complex machines. In the 1600s, the British philosopher Thomas Hobbes suggested that thinking arose from small mechanical motions in the brain. By the 1700s, discoveries about electricity and chemistry led to new theories of human intelligence – again, largely metaphorical in nature. In the mid-1800s, inspired by recent advances in communications, the German physicist Hermann von Helmholtz compared the brain to a telegraph. Each metaphor reflected the most advanced thinking of the era that spawned it. Predictably, just a few years after the dawn of computer technology in the 1940s, the brain was said to operate like a computer, with the role of physical hardware played by the brain itself and our thoughts serving as software. (Epstein 2016)

The hypothesis that our thinking about the mind is characterized by a historical conceptual dependency on pre-existing technological developments[58] is further confirmed when we consider a concept which found its way into cognitive science fairly recently: the distinction between on-line and off-line cognition. This paper will – among other things – further investigate this distinction, as well

[58] Without wanting to underestimate the merits of Zarkadakis' research, it should be noted that his historical observations do not in themselves rule out the possibility that the brain actually is a kind of computer. Zarkadakis' and Epstein's approach certainly 'places things in perspective', but in the final run, the historical approach provides insufficient argumentative means to draw any conclusions regarding the accurateness of the brain-as-computer idea. In any case, trying to decide on this matter will not be an issue for the present article, but see Hutto, Myin, Peeters & Zahnoun *forthcoming* for an elaborated critique of the idea that cognition is computation-based.

as its role in contemporary discussions about the scope of non-representationalist approaches to cognition. For as we'll see, on a standard interpretation, representation-eschewing approaches are accepted as perhaps fruitful when it comes to forms of on-line cognition (coupled perceptuo-motor interactions with what is present in the environment). But when it comes to the higher off-line forms of cognition (decoupled 'mental' activities involving the absent, such as imagining, remembering and reasoning), many are convinced that we need to fall back on an internal- representation-invoking story. In what follows, I shall argue that this wide-spread idea is based on a confusion between the *descriptive* and the *explanatory level*. This confusion already looms large in Clark and Toribio's 1994 *Doing without representing?*, a paper regarded by many as the first and most influential expression of the idea that there are principled boundaries to the new non-representationalist approaches. I shall further argue that the confusion is itself the result of the underlying assumption that a representational explanandum necessarily requires a representational explanans. Crucially, this assumption is flawed. I shall begin, however, by further clarifying the central on-line/off-line distinction.

3.2 On-line/off-line cognition

Historically speaking, the distinction between on-line and off-line cognition appears to have been introduced in cognitive science theory in the first half of the nineties, especially through the work of Andy Clark and his colleagues. In a private e-mail conversation I had with him, Clark says that he himself is unsure how and when exactly it became custom to apply the on-line/off-line distinction to cognition, but he claims that it was "common enough back in the early 90's when Chalmers, Grush and I were all working in the PNP program at Washington University." This would have been around 1993 earliest, as Clark took up a position at Washington University that year as director of the Philosophy/Neuroscience/Psychology program. Of course, the on-line/off-line distinction itself derives from developments in computer technology and

telecommunications, where it refers to a state of either being connected or disconnected to a network, and to the internet in particular. Nowadays, the distinction is familiar enough, which is obviously a direct consequence of the popularization of internet usage. To be sure, it is also in this now popular sense of connectivity versus non-connectivity that Clark and his colleagues started using the on-line/off-line distinction in relation to cognition. In other words, the idea of understanding cognition in computer science terms (which had already been in vogue since the fifties), acquired an 'upgrade' in the nineties thanks to new developments in the same field of computer technology. From then on, cognition couldn't only be grasped in terms of hardware implementing software; now, the mind could also go 'on-line' or 'off-line', depending on the kind of task it is supposed to be performing.

Clark first mentions the distinction between on-line and off-line cognition on paper in his 1996 book *Being There*. It should be noted, however, that in this work, the distinction itself nowhere becomes a topic in its own right. Rather, what we find are more or less *ad hoc* applications of the terms 'on-line' and 'off-line' to distinguish different forms of cognition. The use of the term 'on-line' in relation to cognition is rather ubiquitous in this work, where it is supposed to pick out those forms of cognition that are "intimately dependent on properties of the local environment." (Clark 1996: 63) Examples include visual search, object-identification, tool-use and other intelligent behavior that is said to be 'coupled' to the environment. The term 'off-line', on the other hand, appears only once in the book. In discussing the difference between creatures incapable of internally representing the world versus those that do have this capacity, Clark writes:

> Such creatures are the most obvious representers of their world, and are the ones able to engage in complex imaginings, *off-line* reflection, and counterfactual reasoning. Problems that require such capacities for their solution are *representation-hungry*, in that they seem to cry out for the use of inner systemic features as stand-ins for external states of affairs. (Clark 1996: 147; my italics)

For our purposes, this passage is particularly revealing as here, we see a connection being made between off-line cognition and so-called representation-

hungry problems. As will be explained below, to many working in the field, the distinction between on-line and off-line cognition, as well as the idea of representation-hungry cognition, marks a principled limitation on non-representationalist or radical embodied/enactive approaches to cognition. As Clark remarks a few years later:

> The major challenge for the vision of "radical embodiment" ... lies with the class of "representation-hungry" problems and the phenomena of off-line, abstract, and environmentally decoupled reason. (Clark, 1999, p. 350)

3.3 E-cognition, representation-hunger and radical embodied cognitive science

The diagonal slash dividing on-line from off-line cognition can be seen as emblematic for the dividing line which runs through contemporary cognitive science, separating classic representationalist approaches and more conservative E-approaches from the radical non-representationalist versions known as radical E-Cognition, or REC for short. The prevalent view now seems to be that a non-representation-invoking approach might be warranted or even preferable when it comes to forms of on-line cognition, but when it comes to explaining forms of off-line cognition, however, it is still held by most that we have to fall back on a classic story told in terms of internal representations. To explain the catching of a baseball or the visual tracking of an object, embodied cognition approaches might prove their worth. But when it comes to the heavy lifting, i.e., explaining *real* cognitive phenomena like memory, imagination and thought, matters are best left in the hands of the classic cognitivist. This would mean that non-representationalist E-cognition's explanatory domain is principally restricted in that potential explanations of precisely those forms of cognition that are deemed by many to be genuinely cognitive in nature, fall beyond its scope. Proponents of radical E-cognition, of course, disagree. To better understand the discussion, I will first say a little bit more about E-cognition, radical E-cognition and its relation to classic cognitive science.

The term ‘E-cognition’ covers and unites a number of different, yet often compatible and partially overlapping approaches to cognition which, when taken together, are by many considered to constitute a new paradigm within cognitive science. Usually, four different E-approaches are distinguished: embodied, enactive, embedded, and extended theories of cognition (which is why these approaches also go by the name ‘4E-Cognition’). Perhaps the best way to capture the relative differences between these approaches is to say that they each lay different emphases. Embodied cognition highlights the involvement of the body for cognition; enactive approaches emphasize the ongoing activity of bodily interactions with environments, interactions which are said to bring forth, or *enact*, the world. Embedded approaches stress the importance of the fact that an organism is always already in an environment, and that this embeddedness is always explanatorily relevant. Extended cognition, finally, puts forward the idea that cognition is not restricted to what goes on in brains. It can be co-constituted by environmental elements as well, thereby extending “the bounds of cognition” (Adams & Aizawa 2008). In addition, a fifth ‘e’ could be added when we take into account the significant influence of Gibsonian ecological psychology to contemporary E-cognition. And although it is still too early to speak of E-cognition as *the* new paradigm in cognitive science, the rapidly emerging character of this “new way of thinking about the mind and things mental” (Rowlands 2010: 1) is acknowledged even by its opponents. As classic cognitivist Fred Adams puts it, E-approaches are “sweeping the planet” (Adams 2010a: 619).

Here, I want to focus on an element which we find in some versions of E-theory and which marks perhaps the biggest departure from classic cognitive science: many E-theorists are *in various degrees* skeptical about the explanatory credentials of the notion of internal representation. In its most radical form, known as ‘Radical Embodied Cognitive Science’ (or ‘REC’, for short) internal representations are rejected all-together. The position is nowadays associated with authors like Tony Chemero, Daniel Hutto & Erik Myin, philosophers who have devoted entire books to the explication and propagation of the position’s central tenets (see Chemero 2009; Hutto and Myin 2013; Hutto and Myin 2017),

but the position was already anticipated, and its central thesis already defined by Clark in 1996:

> *Thesis of Radical Embodied Cognition*: Structured, symbolic, representational, and computational views of cognition are mistaken. Embodied cognition is best studied by means of noncomputational and non-representational ideas and explanatory schemes involving, e.g., the tools of Dynamical Systems theory. (Clark 1996: 148)

In other words, radical embodied cognition (henceforth, REC) should not be understood as one more E-approach, but rather as itself an approach to E-approaches. It claims that embodied, enactive, embedded and extended approaches to cognition are best understood in non-representation invoking terms. An often raised question here is: are *all* forms of cognition best understood in a non-representationalist fashion, or only *some*? As indicated above, a common reply is that REC's scope is restricted to only certain forms of cognition. I'll give an illustration of such a reply.

3.4 Will it scale up?

In a famous target article in *Adaptive Behavior*, Randall Beer (2003) presents us with a simulated agent capable of so-called categorical perception, in this case, the successful classification of circle-shaped and diamond-shaped objects in its environment[59]. Beer's model agent "can move horizontally while objects fall from above. It uses an array of seven distance sensors to discriminate between circular and diamond-shaped objects, catching the former while avoiding the latter….The network architecture consists of seven sensory neurons fully connected to five fully interconnected interneurons, which are in turn fully connected to two motor neurons."(Beer 2003: 213). We do not have to go into technical details here. The main point Beer wants to make is that the successful categorizing behavior of his model agent – a dynamically evolving connectionist system – can be described and explained in an entirely representation-eschewing vocabulary, by using tools

[59] Similar model agents already appear in earlier work by Beer. See Beer 1996.

from dynamical systems theory. The explanation is one in terms of the dynamical interactions between the agent and its environment. But at no point must we assume that the agent has constructed internal representations of its environment in order to explain its behavior.

The title of Shimon Edelman's commentary on Beer's target article *But will it scale up? Not without representations* (2003) aptly summarizes the commentary itself. In a passage also quoted by Chemero (2009), Edelman writes:

> Beer's anti-representation stance seems to be unwarranted…in the light of his own example of a system evolved to categorize simple shapes. First, the analytical methods he marshals are barely up to the task even in the toy setting of his choice.... Second, and perhaps more importantly, the target of the analysis — the evolved solution to the toy task — seems to be hardly worth the effort.... All this suggests that drawing wide-reaching conclusions about the nature of, and the need for, representations on the basis of a system encouraged not to have any (by being confronted with but a single toy task) merely breeds doubts concerning the ability of the anti-representation theories to scale up. (Edelman 2003: 274)

Replies such as these are standard, and older examples could be cited. In fact, it seems that ever since radical non-representationalist proposals entered the scene, they have been met with replies serving the double purpose of highlighting the thin application base of non-representationalist approaches, as well as reaffirming the indispensable nature of internal representations for the explanation of 'real' cognition. Since Andy Clark and Josefa Toribio 1994, it has become custom to put the discussion in terms of 'representation-hungry' versus 'non-representation-hungry' problems. The claim is always the same: perhaps REC might be able to deal with the latter kind of problems (problems like the ones with which Beer's model agent is faced), but when it comes to representation-hungry problems, REC's explanatory resources fall short. To see whether this claim is warranted, we should start by asking: What, exactly, are 'representation-hungry problems'?

3.5 Doing without representing? Representation hunger and off-line cognition

As just mentioned, non-representationalist approaches to cognition have been met with skepticism regarding their scope from the very beginning. Probably the earliest, and at the same time most influential expression of this skeptical attitude can be found in Clark and Toribio's *Doing without representing?* (1994). In this paper, the authors critically assess some of the pioneering work of non-representationalist cognitive scientists like Rodney Brooks, Tim van Gelder and the already mentioned Randall Beer. Two central claims against REC can be distinguished. In addition to the claim that REC has insufficient means to deal with 'representation-hungry' problems, the authors contend that the non-representationalist accounts by Brooks, Beer and others are perhaps not as 'representation-free' as these theorists like to think. I will discuss both claims separately, starting with the latter.

Claim A: Non-representationalist accounts are insufficiently non-representational.

Before introducing the idea of representation-hungry problems, Clark and Toribio start their paper by pointing out that, what some theorists want to dub a non-representation invoking explanation of cognition might, on closer inspection, turn out to be undeserving of the label. It is argued that, although the systems of Brooks and Beer do not require *certain kinds* of representations for their explanations, they remain representational nonetheless. Clark and Toribio think that dubbing the kind of explanations Brooks and Beer have to offer 'non-representational' is an overstatement:

> Such overstatement is rooted, we suggest, in an unwarranted conflation of the fully general notion of representation with the vastly more restrictive notions of explicit representation and/or of representations bearing intuitive, familiar contents. (Clark & Toribio 1994: 402)

Clark and Toribio remind the non-representationalist that the key disagreement between classicists and connectionists was not about whether or not internal representations could be dispensed with. The discussion merely concerned the

nature of internal representations, not their existence or explanatory role. Classicists conceived of internal representations as sentence-like strings of symbols which could be manipulated by a certain computational architecture.

> By contrast, connectionists opted for an architecture in which representation and processing were deeply intertwined, and strings of symbols participating in 'cut and paste' processing were replaced by episodes of vector to vector transformation in high dimensional state spaces. (Clark & Toribio 1994: 403)

By overlooking the distinction between these two kinds of representation – a distinction put in terms of explicit vs. implicit representation – theorists like Brooks and Beer are claimed to vastly overstate their case when they present themselves as offering representation-eschewing explanations. Perhaps they do not rely on the explicit language-like representations advanced by the classicist, but their systems remain prone to representational explanation nonetheless. The reason is that, according to Clark and Toribio, Brooks and Beer's accounts can still be counted as connectionist in nature and can therefore be seen as "falling into a more generally representationalist camp." (Clark & Toribio 1994: 412)

Claim B: Non-representationalism ends where representation-hungry problems begin.

After having discussed the allegedly non-representational accounts of Brooks and Beer, Clark and Toribio conclude that, despite their potential merits,

> none of this, on the face of it, amounts to much in the way of evidence for what we shall now dub the General Radical Claim, viz. the claim that internal representation is not essential to genuine cognition. (Clark and Toribio 1994: 412)

The authors go on to discuss another landmark attempt at providing a non-representational explanation of cognitive behavior, the by now well-known dynamical system's analysis of the Watt Governor by Tim van Gelder (van Gelder

1995)[60], which we've already discussed at length in the first chapter. But in the final analysis, Clark and Toribio conclude that, also in van Gelder's case,

> [t]he basic trouble is one that afflicts all the case studies mentioned above. It is that the kinds of problem-domain invoked are just not sufficiently 'representation hungry'. (Clark and Toribio 1994: 418)

Clark and Toribio define the idea of a 'representation-hungry' problem domain as follows:

> By a 'representation-hungry' problem domain we mean any domain in which one or both of the following conditions apply:
>
> 1.The problem involves reasoning about absent, non-existent, or counterfactual states of affairs.
>
> 2.The problem requires the agent to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly (for example, open-endedly disjunctive). (Clark and Toribio 1994: 419)

The subdivision of the representation-hungry problem domain is sometimes grasped in terms of cognition relating to 'the absent' on the one hand, and to 'the abstract' on the other (Clark 1996; Degenaar & Myin 2014). As indicated by Clark and Toribio, the first subdomain involves the absent in three different ways (spatiotemporal absence, non-existence and counter-factuality). In addition to reasoning, it seems sensible, then, to not only include this cognitive capacity, but other forms of cognition that involve the absent as well, in particular memory and imagination.

The second subdomain is said to involve the abstract in the sense that certain problems require sensitivity only to certain salient features of the sensory array. Otherwise put, these are problems that require the cognitive capacity to abstract the same significant features away from different sensory inputs. Clark and Toribio mention as an example the ability to respond selectively "to all and only

[60] At the time of Clark and Toribio's 1994 paper, van Gelder's much discussed 1995 'What might cognition be if not computation?' was still in press.

those items which belong to the Pope" (Clark & Toribio 1994: 420), although 'edibility' would probably be a less far-fetched example.

The point Clark and Toribio want to bring across is clear enough: cognition involving the absent and the abstract is 'representation-hungry' in that it requires internal representations that stand in for both absent and abstract properties. Furthermore, since representation-hungry cognition is the kind of cognition that is deemed 'genuine', it follows that genuine cognition requires invoking internal representations as well. So whatever it is the non-representationalist thinks he has achieved, for Clark and Toribio, it is clear that it is not to be considered as an explanation of 'genuine' cognition.

Note that, when we view the discussion in light of the distinction between on-line and off-line cognition, it appears that this distinction coincides[61] with Clark and Toribio's distinction between representation-hungry and non-representation-hungry cognition, and, consequently, with the distinction between genuine and non-genuine cognition. As we've seen, off-line forms of cognition pick out decoupled forms of cognition that deal with the *absent* (memory, imagination…) as well as the *abstract* (reasoning, categorization…), and are therefore said to require internal representations, whereas 'on-line cognition' (perceptuo-motor interactions) picks out coupled cognitive activity that deals with what is *present* here-and-now, and with what is *concrete* instead of abstract. On Clark and Toribio's account, it becomes questionable, then, whether on-line cognition even qualifies as a form of cognition[62] at all.

In the following, I want to critically assess Clark and Toribio's two central claims. As I will argue, both the claim that the discussed non-representationalist accounts are still representational (Claim A), as well as the claim that non-

[61] The conceptual correspondence between representation-hungry cognition and off-line cognition is confirmed by Clark himself. See Clark 2005.

[62] Clark and Toribio 1994 certainly lean towards this idea. However, the dismissal of on-line activity as a form of cognition is much more straightforward with cognitivists like Fred Adams, Kenneth Aizawa and many others defending the idea that the involvement of representations is a *conditio sine qua non* for cognition. For these authors, the involvement of internal representations constitutes "the mark of the cognitive" (for recent accounts, see Adams and Aizawa 2008, Adams 2010b, Adams and Garrison 2013). To the extent, then, that on-line cognition can be understood without invoking representations, it follows that this form of cognition shouldn't even qualify as properly cognitive to begin with.

representationalist approaches are confronted with a principled barrier when it comes to off-line cognition (Claim B) are unwarranted. Notably, because of the second claim's centrality for our discussion, I will devote substantially more attention to the critical assessment of claim B.

3.5.1 Against claim A: Ramsey's 'job description challenge'

As we've seen, Clark and Toribio are unconvinced by the idea that the self-proclaimed non-representational approaches by Brooks and Beer can be properly labeled as such. By pointing out the close affiliation of the latter's allegedly non-representation involving designs with connectionist modelling, the authors suggest that the systems developed by Brooks and Beer are still representational in the way that connectionist systems are said to be representational. Clark and Toribio's argument, then, hinges on the claim that connectionist systems are in fact representational systems. As is by now clear, this claim is not well-founded. In the past decade, authors like William Ramsey have tried to show how the idea of characterizing connectionist systems as representation-using systems is based, not so much on theoretical or empirical grounds, but rather on a habit of mind. Ramsey's 2007 book *Representation Reconsidered* starts with the critical observation that, within cognitive science, there is "an excessive over-application" of the notion of representation (Ramsey 2007: i). Ramsey closely examines the various ways in which the notion of representation is being used in contemporary cognitive science and puts them up against, what he calls, the 'job description challenge'. This is "the challenge of explaining how a physical state actually fulfills the role of representing in physical or computational processes – accounting for the way something actually serves as a representation in a cognitive system." (Ramsey 2007: xv) Spelled out in a little more detail:

> There needs to be some unique role or set of causal relations that warrants our saying some structure or state serves a representational function. These roles and relations should enable us to distinguish the representational from the non-representational and should provide us with conditions that delineate the sort of

> job representations perform, qua representations, in a physical system. (Ramsey 2007: 27)

Ramsey concludes that, when it comes to the way in which the notion of representation is used in classical computational theories, the job description is satisfactorily being met. However, when it comes to newer approaches, including connectionism, it is not: "Although neuroscientific and connectionist theories characterize states and structures as inner representations, there is, on closer inspection, no compelling basis for this characterization." (Ramsey 2007: xiii) Cast in terms of Clark and Toribio's distinction between explicit and implicit representations, Ramsey argues that the postulated implicit representations we find in connectionist accounts are not doing anything recognizably representational at all, and we have therefore no reason whatsoever to still think of these systems as representational. Ramsey provides us in addition with an explanation of why the notion of representation came to get so easily adopted by connectionism, even though it is explanatorily superfluous:

> When new scientific theories are offered as alternatives to more established views, proponents of the new perspective are sometimes reluctant to abandon the familiar notions of the older framework, even when those posits have no real explanatory role in the new accounts. When this happens, the old notions may be re-worked as theorists contrive to fit them into an explanatory framework for which they are ill-suited. (Ramsey 2007: xiv)

So although some cognitive scientists operating within the connectionist framework are still hanging on to the idea of internal representation, the notion is actually no longer doing any explanatory work and may just as well be abandoned. Clark and Toribio's claim that the accounts by Brooks and Beer are still representational in the sense that they rely on implicit representations therefore appears, in light of the above, unwarranted. Connectionist accounts shouldn't be understood as representational in the first place.

3.5.2 Against claim B: a fundamental fallacy

Before starting my rather extensive argument against Clark and Toribio's second claim, namely that non-representational accounts end where genuine cognition begins, I want to again stress that the idea that the existence of representation-hungry cognition imposes principled limits on non-representationalist approaches, is widely accepted. Notably, the idea has found its way into the Stanford Encyclopedia of Philosophy, where it isn't merely being explained, but at the same time endorsed. In their entry on embodied cognition, Robert Wilson and Lucia Foglia write:

> Formulating an empirically adequate theory of intelligent behavior without appealing to representations at all … faces *insuperable* difficulties…For example, organism-environment interaction alone cannot account for anticipatory behavior, which involves internal factors beyond the immediate constraints of the environment to achieve or fulfill future needs, goals or conditions. Domains raising a representation-hungry problem (…) are those involving reasoning about absent, non-existent or counterfactual states of affairs, planning, imaging and interacting (Wilson and Foglia 2011: section 4.2, my emphasis)

Consequently, trying to argue against the claim that representation-hungry cognition falls outside the scope of non-representationalist cognitive science is doing something more than arguing against Clark and Toribio alone. It means going against a view which has arguably become the standard view when it comes to the reach of the non-representationalist outlook.

To be sure, arguing against the idea that representation-hungry cognition poses principled limits on non-representational approaches can be done in more than one way. As Chemero notes, one "possibility is to use empirical work to show that radical embodied cognitive science has the resources to explain representation-hungry tasks." (Chemero 2009: 40)[63] However, "showing by example that there is no in-principle reason that radical embodied cognitive science is not capable of explaining "real cognition"" (Chemero 2009: 42-43) is not the kind of strategy I

[63] Chemero refers to work by Van Rooij, Bongers, and Haselager. See Van Rooij, Bongers and Haselager 2002.

want to pursue here. The reason is that this would already grant Clark and Toribio too much. Their claim does not need to be countered with empirical evidence, because, on closer examination, it turns out that it is not clear what the claim is, exactly. Underneath its surface, Clark and Toribio's 'Argument from Representation-Hunger' (henceforth ARH) is at once fundamentally ambiguous, analytical, as well as stipulative. More importantly, it is also unscientific in that its claim is grounded, not in any empirical considerations, but in a flawed assumption, namely the assumption that representational cognition requires representational explanations. The following analysis serves to substantiate these charges.

3.6 Ambiguity with regard to description/explanation

The first thing that needs emphasizing is that the notion of representation-hunger is fundamentally ambiguous with regard to the distinction between descriptions and explanations. Throughout Clark and Toribio's paper, the qualification 'representation-hungry' appears to be serving the double purpose of descriptively picking out a certain kind of cognition – which the authors equate with 'genuine' cognition – as well as characterizing the kind of explanations this 'genuine' cognition requires. Elsewhere in the literature, we encounter the same ambiguity in expressions like 'representation-*involving*', or '*representational* cognition', expressions which have become idiomatic within philosophy of cognitive science, but which nonetheless leave unspecified whether one is referring to cognition on a descriptive, rather than an explanatory level. Recall Wilson and Foglia's definition of representation-hungry problems as "those involving reasoning about absent, non-existent or counterfactual states of affairs, planning, imaging and interacting". The problem is that formulations in terms of 'the involvement' of representations leave unspecified whether representations are involved only *descriptively* or also *explanatorily*. What we often find, however, is that these formulations are supposed to pertain to both descriptions and explanations, and that the distinction between the descriptive and the

explanatory level is simply ignored. Clark and Toribio's notion of representation-hunger proves no exception to this. Is it a certain form of cognition (i.e., off-line cognition like memory, imagination and reasoning) that is said to be 'representation-hungry', or does the notion apply to the explanations of these forms of cognition? In other words, does 'representation-hungry' apply to the explanandum or the explanans?

As said, the answer is that it applies to both. On the one hand, it is rather obvious that the idea of representation-hunger must pertain to *explanations* of cognition. How else is the idea supposed to be relevant for discussions about the *explanatory* scope of non-representationalist approaches? So clearly, ARH wants to be an argument at the explanatory level. Therefore, it is the *explanation* of off-line cognition that must be deemed representation-hungry in the sense of requiring the positing of internal representations. On the other hand, it is equally obvious that representation-hunger applies, not only to the explanations, but at the same time also to the kinds of cognition under consideration. It is no coincidence that 'representation-hungry cognition' covers precisely those forms of cognition that are commonly characterized as representational in nature. Within phenomenology, for instance, memory, imagination and other forms of off-line cognition are canonically described as representational because of their intrinsic intentional relation to what is not-present. As we'll see, 'off-line cognition' captures precisely the kind of mental activity that both classic, as well as contemporary phenomenologists refer to in terms of mental representation. Unlike perception, cognitive activities like remembering or imagining are best *described* as representational, in that they somehow re-present that which is in some sense absent, or at least not present in perception. I'll return to this. What I want to argue first of all, however, is that the ARH poses no threat to the non-representationalist, regardless of whether we consider 'representation-hunger' at the level of description, or at the level of explanation.

3.6.1 'Representation-hunger' as a descriptive notion

To better see why, at the descriptive level, the idea of representation-hunger is no problem for the non-representationalist, consider a random memory-task, like remembering what you had for breakfast this morning. Authors like Clark and Toribio will agree that this is a case involving a kind of cognition to which qualifications like 'off-line', 'representation-hungry' and 'genuine' surely apply. However, the only reason we have for calling such a task 'representation-hungry' is that it by definition involves memory, something which is *already understood as a representational capacity*. We may just as well call a cognitive task like recalling what one had for breakfast 'memory-hungry' instead of 'representation-hungry'. It is only because we have already agreed on defining the activity of remembering in terms of *re-presenting*, for instance a past episode, that the idea of representation-hunger seems appropriate. But note that this makes the idea of labelling memory-involving cognition 'representation-hungry' completely analytical. If the act of remembering is already defined as an act of representing, and if an act of representing is already defined as necessarily involving mental representations, then of course, any problem that is describable as involving memory, or memory-hungry, is also describable as representation-hungry. But this is an entirely analytical, and empirically uninteresting idea. More importantly, however, nothing about this idea forms a threat to the non-representationalist program of providing *explanations* of cognition without positing internal representations. We are after all still on the descriptive level. One can easily acknowledge that a certain cognitive phenomenon is best described as representational in nature, without therefore having to accept the very different idea that these phenomena need to be explained in terms of the processing of internal representations. Even if we concede that the various forms of off-line cognition are best described as representational capacities, it doesn't follow that we need to *explain* these capacities by invoking internal representations. By analogy, one might very well accept the idea that the heart is best described as a pump, without therefore having to concede that, somewhere in the explanation of the organ's function, we will have to invoke the notion of (internal) pumps. In fact, thinking that we *must* seems like a particularly bad idea

as it immediately raises worries of circularity or regress. So why, then, should we so easily accept the idea that cognition that is best characterized as representational *must* be explained by invoking internal representations? I'll suggest an answer to this question in the following section, and, in more detail, in section 3.12.

3.6.2 'Representation-hunger' as an explanatory notion

On the explanatory reading, Clark and Toribio's claim is that certain forms of cognition can only be *explained* by invoking internal representations. This is the sort of claim needed to provide a potential threat for the non-representationalist. The problem, however, is that we do not find a single argument for this claim, which is nevertheless the potentially interesting one (the descriptive claim being, as we've just seen, analytical). I want to suggest that the reason why someone might feel that there *is* an argument here is because Clark and Toribio appeal to, and invigorate an implicit assumption or intuition which appears to be well-entrenched within mainstream cognitive science, an intuition which can be formulated as the idea that *whatever explains our representational capacities (as exhibited in memory, imagination and thought) must be sufficiently similar to it* (I'll bring this point back up in section 3.4). Ultimately, all Clark and Toribio are doing is rehearsing the unwarranted assumption that *representational cognition must be explained by representational explanations*. This assumption pervades much of mainstream representationalist cognitive science, but it's influence on contemporary theorizing about cognition is, as I will hope to show below, perhaps nowhere as apparent as within prominent empirical research on imagination and imagery. Before I turn to this, however, I want to end my discussion of Clark and Toribio's notion of representation-hunger with a methodological consideration pertaining to their idea that only representation-hungry cognition counts as *genuine* cognition (which, of course, only further devaluates potential non-representationalist approaches). As has recently been

argued, postulating the involvement of inner representations as a 'mark of the cognitive' is probably not a good idea.

3.7 Real cognition as representation-hungry cognition

As we've seen, in their paper, Clark and Toribio defend the idea that representation-involving cognition requires representation-involving explanations and that, in addition, only this type of cognition (off-line cognition) can be said to be "genuine cognition", precisely *because* of their representational nature. As I've already pointed out, Clark and Toribio do not give an actual argument for the former assertion of real cognition requiring representational explanations. But neither do they present us with an argument for the latter claim that only representation-hungry cognition qualifies as genuine cognition. Although this latter idea is almost as widely-accepted as the former, it is ultimately no more than a stipulation. We might wonder, however, whether this specific demarcation of genuine cognition in terms of representation is, from a methodological perspective, such a promising idea. William Ramsey, for one, argues convincingly that it is not. In a recent paper of the same title, Ramsey asks himself: "Must cognition be representational?" (Ramsey 2017). The central focus of the paper precisely lies with the demarcating role of representations. His main claim is:

> Even if you think cognitive scientists must invoke representations to explain a wide array of cognitive capacities and processes, I'll argue you should nevertheless reject the use of representations to define cognitive processes and theories.... My goal here is to establish not an anti-representational thesis, but rather an anti-representation-as-definer-of-cognition thesis. (Ramsey 2017: 4198)

Ramsey substantiates this latter thesis with three arguments. Defining genuine cognition in terms of representation should be rejected because, first, "it puts undue restrictions on psychological theorizing" (Ramsey 2017: 4201); second, "it undermines the representational theory of mind" (Ramsey 2017: 4205); third, "it encourages wildly deflationary accounts of representation" (Ramsey 2007: 4206).

All three arguments contain elements that are relevant to our topic, i.e., the scope of non-representationalist cognitive science. I shall therefore briefly discuss each of them in turn.

Argument 1: defining cognition in terms of representation puts undue restrictions on psychological theorizing

By placing a priori constraints on the kinds of theory that are eligible as theories of cognition, we run the risk of excluding perhaps unexpected, though potentially ground-breaking original approaches to cognition. Not only has the history of science shown us that scientific breakthroughs are achieved precisely by giving up on certain theoretical frames, the idea of defining cognition in terms of representations is itself not warranted by scientific research. It is something that largely stems from our pre-scientific common sense notions of the mental. As Ramsey puts it, "it seems quite clear that representationalism is a core dimension of our ordinary, common-sense or "folk" psychology." (Ramsey 2017: 4202) The essential point of the first argument, then, is

> …that there is at least some reason to think our tendency to *define* cognition in representational terms or to simply equate mental processes with representational processes stems from presuppositions that are not well-founded and that are prone to hinder scientific theorizing. (Ramsey 2017: 4202-3)

Argument 2: defining cognition in terms of representation undermines the representational theory of mind

Ramsey's second argument is particularly relevant for our discussion and comes very close to what I have been arguing above. Earlier, I raised the concern that the idea of applying the qualification 'representation-hungry' to both explanandum and explanans is, epistemologically speaking, probably a bad idea. In his second argument, Ramsey further spells out this idea. The critical questions he raises here are important, and I'll quote them at length:

> [H]ow can representationalism be *both* a falsifiable explanatory theory about the building blocks of cognition, and, at the same time, an essential, defining criterion of cognition itself? How can the positing of inner representations be an

> explanatorily novel and illuminating feature of theories that are supposed to account for various cognitive processes, and simultaneously be a necessary, qualifying condition for *theorizing about* cognitive processes? …How did we come to treat an allegedly empirical hypothesis about mental phenomena as a way of defining mental phenomena? (Ramsey 2017: 4205)

As Ramsey further correctly points out, "[y]ou can't treat representational posits as *both* interesting explanatory constructs *and* as a necessary condition for a legitimate account of the phenomena you are trying to explain." Indeed. Yet, the fact that this is precisely what is going on in mainstream cognitive science (and in accounts like Clark and Toribio's) bears testimony to the kind of confusion I've been trying to expose in the above, namely a confusion, or better, conflation of the descriptive with the explanatory. Again, the idea that genuine cognition must be explained by invoking representations seems to be undergirded, not by scientific research, but by the wide-spread, yet confused assumption that cognitive phenomena that are representational in nature can only be accounted for by explanations that are also representational in nature. As already mentioned, there seems to be precisely no reason why we should accept this assumption, as it directly raises issues of potential circularity or regress.

Argument 3: defining cognition in terms of representation encourages wildly deflationary accounts of representation

Ramsey's third argument reconnects with Clark and Toribio's first claim. To recapitulate: Clark and Toribio argue that certain theorists like Brooks and Beer are somewhat overhasty in labelling their account 'non-representationalist'. Their models do not perhaps appeal to the explicit representations we find in classical computational accounts of cognition, but that doesn't mean that they can't be identified as representational in a different, more general sense, for instance in the sense that connectionism can be said to be representational. I've relied on Ramsey 2007 to refute the claim that the notion of representation we find in connectionist theories is doing anything recognizably representational in nature. Ramsey 2017 returns to this point. According to him, the fact that we find such "wildly deflationary" notions of representation within cognitive science is itself

encouraged by the common practice of defining cognition in terms of representation. Apparently, it is felt that without representations, a theory (e.g., a connectionist theory) simply doesn't qualify as a theory of *cognition*. Ramsey quotes philosopher Murat Aydede here, who, in discussing connectionism, claims that it is difficult to see how connectionism can be said to be modeling *cognitive* phenomena if the units of the networks aren't treated as representing[64]. Since it is not clear, then, how these units can be said to be representational in any classical sense, they are simply taken to be representational in a different sense, resulting in an unwarranted deflationary notion of representation. It isn't much of an overstatement, then, to claim that mainstream cognitive science is characterized by a climate of representation-despotism, a climate in which apparently a theorist must first show how his or her hypothesis is sufficiently representation-reliant in order for the theory to be even considered as a potential theory of cognition. No wonder, then, that non-representationalist approaches are met with such resistance.

3.8 More ambiguity: 'representational' as a predicate

In light of our discussion of the scope of non-representationalist approaches to cognition, I would like to retain the two following elements from Ramsey's arguments. First, the observation that the idea that cognition must be defined in representational terms is not only unscientific (in the sense that it is not grounded in scientific research), it is anti-scientific (in the sense that it hinders scientific progress). Second, the observation that mainstream cognitive science is characterized by a confusion between description and explanation or explanandum and explanans. Yet, despite Ramsey's important and insightful criticisms, his argument (including the main question of his paper, "Must cognition be representational?") is itself still characterized by an ambiguity. The term 'representational' is, in Ramsey's account as elsewhere, ambiguous in that it

[64] See Ramsey 2017: 4207. See also Ramsey's reference to Aydede: Aydede, M. (2010). The language of thought hypothesis. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (fall 2010 edition). Retrieved from http://plato.stanford.edu/archives/fall2010/entries/ language-thought/.

isn't clear whether it is supposed to pertain to the nature of cognition, or to the nature of the explanations of cognition. As with Clark and Toribio, the answer seems to be "both". Throughout his 2017 paper, Ramsey's discussion pertains to how cognition is being defined, as well as to how potential explanations of cognition are being delimited. Oddly enough, Ramsey decides to compile claims pertaining to the nature of cognition with claims pertaining to the nature of the explanations of cognition. Both kinds of claims (descriptive and explanatory claims) are included in what Ramsey calls the "representational demarcation thesis" (Ramsey 2017: 4200). This is understandable in that he wants to say something about the way the notion of representation serves as a criterion for genuine cognition, as well as genuine explanations of cognition. However, this merger has the undesired effect of making the term 'representational' seem more univocal than it really is. Ramsey is surely right that, within mainstream cognitive science, the notion of representation is ubiquitous on both the descriptive and the explanatory level. He nevertheless fails to point out that the 'representational' attribution – as in "Must cognition be representational?" – *means something very different*, depending on whether we are using it to qualify the cognitive explanandum, or whether we want to characterize its potential explanation. The term 'representational' may mean different things, depending on whether it is attributed to a cognitive phenomenon or a cognitive explanation. This second kind of ambiguity needs addressing as well if we want to get clear on what it means, precisely, to be a non-representationalist with regard to cognition.

3.9 Representational cognition vs. representational explanation in cognitive science theory

As the above analysis should have made clear, mainstream cognitive science can be said to be 'representational' in that it is characterized by 'representational' commitments, both with regard to what counts as a genuine explanation of cognition, as well as to what counts as genuine cognition itself. We see this double commitment returning in the so-called Representational Theory of Mind, or RTM for short. Although its name suggests otherwise, RTM is best understood,

not so much as a well-defined theory, but rather as a generic notion, simultaneously covering claims about the nature of cognition, as well as claims about its explanation. These different claims (i.e., claims pertaining to the nature of cognition and claims about the explanation of cognition) are all part of the RTM, then, in that they are all said to be 'representational'. The problem, however, is that the term 'representation', as well as its cognates 'representational' and 'represent' have become such all-purpose notions that they tend to obscure the fact that these terms mean very different things, depending on whether we are talking about the nature of cognition or about the nature of the explanation of cognition. In the past, this has led to confusions on both sides of the representationalist/non-representationalist divide. Cognitivists, for instance, like to point out that representationalist accounts have a long and rich history, and that the RTM does not begin with Fodor, but goes all the way back to antiquity, at least to Aristotle[65]. Indeed, in his *De Anima*, Aristotle was perhaps the first to *describe* perception as somehow mediated (though not by neural structures), or to *characterize* imagination in terms of inner images. In some sense, then, his account can be said to be 'representational'. But this is a long stretch from the 'representational' *explanations* we find in classic computationalism. For the computationalist, 'representation' refers first and foremost to hypothetical internal content-carrying objects (presumably neurally implemented) over which computational operations are defined, and which have an indispensable explanatory role within a causal account of cognition. So although there is probably some historical kinship between Aristotelean psychology and modern day cognitivist approaches to cognition, the idea of postulating representations as internal objects to be operationalized within computational explanations of cognition did, of course, not occur to Aristotle.

A different and more actual example of the 'representational description/explanation' conflation can be found in Lawrence Shapiro's review of Hutto & Myin's *Radicalizing Enactivism*. As already mentioned, Hutto & Myin's

[65] See for instance David Pitt's entry on Mental Representation in the Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/mental-representation, section 1. See also Ramsey 2017: 4202.

work is dedicated to a defense of non-representationalism and to an argument that shows how internal representation-invoking explanations of cognition fall short. To be clear, their argument is aimed at the *explanatory* representational content-invoking framework which has come to dominate the contemporary study of cognition. Shapiro, however, finds Hutto & Myin's project, not so much unconvincing, but rather "inconceivable" (Shapiro 2014: 213). He writes:

> They [Hutto & Myin] deny what seems undeniable—that mental states, like my thought that Madison sits on an isthmus, and even perceptual states, such as my visual experience of sailboats on Lake Mendota, represent anything. But how can a thought that p not be about p? (Shapiro 2014: 213; my addition)

Shapiro's critique, however, is beside the point. By conflating the level of description with that of explanation, he is led to the idea that Hutto & Myin's non-representationalism entails that human thought, for instance, can no longer be *characterized* as representing how things stand with the world. This is *not* what these and other non-representationalist authors are saying. In fact, how could it be? Thinking is indeed defined intentionally (i.e., as being about something), and is in that sense also definable as representational. But such an analytical truth is uninteresting, and it is certainly not the ambition of non-representationalist theorists to argue against things which are true by definition. Proponents of REC have no quarrel with the idea that 'the thought that *p*' has the content *p*. They are not absolute content-eliminativists. What authors like Hutto & Myin *are* arguing against is the idea that cognitive activity of various kinds must necessarily be *explained* by invoking internal content-carrying states as *explanatory* posits. By failing to make the distinction between representational descriptions and explanations, Shapiro's review misses its mark entirely. Furthermore, we might wonder whether Shapiro's (and many others') idea that perceptual activity is representational, and thus content-involving, isn't based on a confusion between the actual perceptual activities, and our descriptions of these perceivings. The phrase 'my visual experience of sailboats on Lake Mendota' can be redescribed as 'my seeing that there are sailboats on Lake Mendota'. The latter is, of course, a content involving description. But we should be wary not to

confuse our content-involving descriptions of our perceptual experiences with those experiences themselves.

3.10 Representational cognition vs. representational explanation in phenomenology

As said, lumping together 'representational' descriptions with internal representation-invoking explanations is something we find, not only in the representationalist camp, but in the anti-representationalist division as well. A good example can be found in the work of Hubert Dreyfus, without doubt one of the most longstanding and staunchest opponents of computational-representational approaches to cognition. Turning our attention to Dreyfus' example will at once allow us to connect our discussion to the topic of mental representation within phenomenology. The point of this connection will be to make as explicit as possible the very different sense in which 'representational' can be predicated, depending on whether we are describing the nature of cognitive phenomena (and especially off-line cognition) or whether we are offering potential explanations of these phenomena. And since the meticulous *description* of the representational character of mental activities like remembering and imagining is the strong suit of Husserlian phenomenology, we might expect some valuable insights here regarding our subject of mental representation as a descriptive notion vs. mental representation as an explanatory posit.

In a much discussed section of his introduction to *Husserl, Intentionality and Cognitive Science* (Dreyfus 1982), Dreyfus maintains that Husserl should be regarded as the father of contemporary representationalist cognitive science. He writes:

> Husserl has finally begun to be recognized as the precursor of current interest in intentionality - the first to have a general theory of the role of mental representations in the philosophy of language and mind. As the first thinker to put directedness of mental representations at the center of his philosophy, he is

> also beginning to emerge as the father of current research in cognitive psychology and artificial intelligence. (Dreyfus 1982: 2)

Dreyfus' depiction of Husserl as "the author of a proto-Fodorian theory of mental representations" (McIntyre 1986: 101) did not escape the attention of Husserl scholars, who vigorously rejected the comparison, and justifiably so. Authors like Ronald McIntyre and, especially, Beth Preston have convincingly shown why it would be a mistake to liken Husserl's phenomenological investigations (which indeed involve mental representation *in some sense*) to the representation invoking explanations that have become so ubiquitous within standard cognitive science. McIntyre and Preston give us a number of reasons for disagreeing with Dreyfus' comparison, two of which are particularly relevant for our discussion. These two reasons are succinctly captured in the final sentence of McIntyre 1986:

> [P]henomenology remained for Husserl a *descriptive* discipline, descriptive of intrinsically *intentional* experiences, as they are *experienced*. (McIntyre 1986: 112; final emphasis mine)

3.10.1 Phenomenological description vs. cognitivist explanation

First, then, Dreyfus' portrayal of Husserl as a kind of precursor to contemporary computationalist cognitive science is unwarranted as it appears to neglect the fact that Husserl's phenomenological enterprise was first and foremost a *descriptive* one. This is not a mere contingency of Husserl's approach, but the essence of the phenomenological method itself. As McIntyre points out, Husserl simply *cannot* be concerned with the explanations of mental states "in terms of their causal relations to one another and to the world, for causality (in any naturalistic sense) is "bracketed" by phenomenological epoché." (McIntyre 1986: 104) On Husserl's account, mental representations are not to be understood as explanatory posits, which is, of course, precisely how they *are* conceived of within computational explanations of cognition. Beth Preston similarly criticizes Dreyfus for conflating Husserl's descriptive analyses with the explanatory

accounts we nowadays find in cognitive science. With regard to the issues of mental content and the intentional nature of the mental, she writes:

> [I]nsofar as Husserl talks about noematic structures of consciousness he is advancing a purely *descriptive* theory of the intentionality of the mental, not an explanatory theory. …Husserl takes meanings as the basic furniture of mental life and wants only to describe the underlying meaning-structure of experience….The RTM theorist, on the other hand, wants to understand how it is that mental symbols come to have the meanings they have, and therefore takes mental meaning as something for which an explanatory account needs to be provided. (Preston 1994: 220)

It appears, therefore, that Dreyfus' interpretation of Husserl has fallen prey to the same confusion we've encountered, for instance in Clark and Toribio's account of representation-hunger, namely conflating descriptions with explanations.

3.10.2 Personal vs. sub-personal mental representation

As said, the second reason why Husserl's phenomenological descriptions can hardly be considered as recuperable within the framework of contemporary cognitive science is already being mentioned in McIntyre's short quote above, but we also find it in the just cited passage from Preston. Perhaps the most obvious reason why we shouldn't think of Husserlian mental representation as being on a par with the kind of internal representations posited by cognitive science is that, on Husserl's account, mental representation is inseparable from conscious experience. Mental representation refers first and foremost to certain subjectively experienced acts of consciousness, rather than to sub-personal object-like entities underlying, and supposedly explaining, our experience. Within mainstream cognitive science, the standard view is that conscious subjective experience is to be explained in terms of the underlying processing of content-carrying mental representations. Here, 'mental representation' picks out first of all the elements (content-carrying vehicles) over which these underlying processes are defined. A good illustration of how this works can be found in the work of Stephen M.

Kosslyn, without doubt one of the leading authors in the scientific study of mental imagery and imagination. Taking a look at Kosslyn's work provides at once a better insight into the scientific study of at least one form of off-line cognition.

Within the context of Kosslyn's extensive research – which now spans over four decades – 'mental representation' does not so much refer to the phenomena under investigation (mental imagery and imagination), but to the sub-personal processes which are said to underlie and explain these phenomena *qua* consciously experienced. The actual experience of imagery is referred to as only "the tip of the iceberg". It follows, then, that phenomenological introspection can only provide but a very limited understanding of mental imagery. As Kosslyn makes clear: "Introspection by its very nature only exposes the tip of the iceberg." (Kosslyn 1978: 225) So although the tip of the iceberg can be characterized in terms of mental representation, it is ultimately the sub-personal and non-experiential mental representations we find in the submerged portions of the "mental iceberg" (Kosslyn 1980: 21) which are of interest to anyone attempting to scientifically explain mental imagery and imagination *qua* experienced phenomena. I will return to this 'two-storey story'[66] in which mental representations are located on both the subjective experiential, as well as the sub-personal, sub-experiential level. For now, I only want to emphasize this second important disagreement between mental representation in Husserl's phenomenological sense, which is always experiential, and the sub-personal, sub-experiential representations as hypothesized by leading cognitive scientists like Kosslyn.

3.10.3 representational capacity vs. representational object

In addition to these two dissimilarities between phenomenological accounts of mental representation and those we find in cognitive science, there is a third significant difference which has up till now remained implicit. Next to being

[66] The term is a variation on Hutto & Myin's "multi-storey story". See Hutto & Myin 2017: 137.

descriptive (instead of explanatory), and experiential (instead of sub-personal), within the phenomenological tradition, the term 'mental representation' refers first of all to a (mental) activity or an 'act of consciousness', and not to some physically implemented intracranial structure or process. 'Mental representation' is something we *do*, rather than some-thing in our head. Or, more precisely, it is something we are *able* to do, not some-thing that *enables* us to do something. Here, mental representation refers to a kind of capacity, and in particular, the capacity involved in acts like remembering, imagining and other conscious activities: the capacity to relate to the absent. As Husserl scholar Eduard Marbach informs us at the beginning of his phenomenological study *Mental Representation and Consciousness*:

> According to Husserl's clarification of the nature of these conscious mental activities…they are all modifications of the less complex mental activity of perceiving. The experience of perceiving something is a mental activity of intentionally referring to something in its *present* givenness, i.e. an activity of presenting something. The experiences of imagining, viewing pictures, and remembering, on the other hand, are so many ways of intentionally referring to something *absent,* i.e. activities of re-presenting something. (Marbach 1993: 1)

Indeed, as already indicated above, what Husserl and other phenomenologists refer to as mental representation appears to pick out exactly those forms of cognitive activity that Clark and Toribio refer to as representation-hungry and which also fall under the heading 'off-line cognition'. As to the nature of the internal representations that mainstream cognitive scientists hypothesize to be underlying, as well as explaining these forms of off-line cognition, these are first of all conceived of as object-like entities with a certain format. As Kosslyn makes clear on several occasions[67], the notorious imagery debate (which started in the mid- seventies) is a debate about *the format* of internal representations, not about whether they exist or not. To this day, much philosophical ink is spilled on the question whether internal representations are best conceived of as linguistic symbol-like, or as picture-like, or as a kind of internal model or map. But that

[67] See, for a more recent example, Kosslyn et al. 2010: 19.

there are such presumably neurally implemented content-carrying entities in the head is a hard-wired assumption of mainstream cognitive science, where cognition has become virtually synonymous with internal-representation-processing. So, unlike phenomenological accounts, here, representations are thought of, not in terms of mental activity, but in terms of *physical objects*, which are conceptually modelled on more familiar external 'representational' objects (words, sentences, pictures, models, maps…). As Dennett already notes in 1978:

> Whatever *mental* representations are, they must be understood by analogy to *non*mental representations, such as words, sentences, maps, graphs, pictures, charts, statues, telegrams, etc. The question is whether any of one's mental representations are more like *pictures* or *maps* than like *sentences,* to take the favorite alternative" (Dennett 1978: 175).[68]

"The favorite alternative", here referring to the just mentioned imagery debate which was, and still is, a debate between those who think of the format of mental representations as picture-like, and those who think of it as sentence-like[69]. It is not an insignificant detail that Kosslyn is to this day one of the main defenders of the idea that the underlying representations that explain 'genuine' cognition like imagery and imagination are picture-like. Indeed, Kosslyn tells a story in which conscious mental representations that are picture-like[70] are to be explained by underlying, sub-personal mental representations which are themselves picture-like, rather than propositional. I will not go into this right now. My point here is to merely show how, within cognitive science, internal representations are thought of as to some extent analogous to external representational *objects*.

This *reified* picture of representation, which has been already discussed extensively in the second chapter, is not only far removed from Husserlian

[68] Ramsey 2007 makes the same point: "We can't posit representational states to do many of the things they are supposed to do in a theory unless the posit itself is sufficiently similar to the sort of things we pre-theoretically think representations are." (Ramsey 2007: 12)

[69] Within the imagery debate, Zenon Pylyshyn can be seen as the main protagonist of the view that the format of the underlying representations is sentence-like, i.e. propositional. See Pylyshyn 1973.

[70] Visual imagery and imagination are on Kosslyn's account also assumed to be picture-like, as his 'mental scanning' experiments make clear. See, for instance, Kosslyn 2010: 25. For a contemporary critique of the idea of thinking of conscious mental imagery as a kind of internal-picture-viewing, see Thompson 2007: Ch. 10.

phenomenology, it appears to be in direct conflict with it. In a revealing passage, Dan Zahavi has recently pointed out that it wouldn't only be a mistake to associate Husserl with the kind of representationalism that is still dominating cognitive science, it would be misguided to understand Husserlian phenomenology as endorsing *any* kind of representationalism *at all*. As Zahavi emphasizes, Husserl's "turn towards transcendental idealism was partially motivated by his rejection of both representationalism and phenomenalism, and by his efforts to safeguard the objectivity of the world of experience." (Zahavi 2018: 56) Husserl's rejection of representationalism, understood as the idea that our experience of the world is always mediated by internal object-like representations, becomes very clear in his 1915 lecture course *Ausgewählte phänomenologische Probleme*. Paraphrasing Husserl, Zahavi writes:

> [N]othing might seem more natural than to say that the objects I am aware of are outside my consciousness. When my experiences – be they perceptions or other kinds of intentional acts – present me with objects, one must ask how this could happen, and the answer seems straightforward: By means of some representational mediation. The objects of which I am conscious are outside my consciousness, but inside my consciousness, I find representations (pictures and signs) of these objects, and it is these internal objects that enable me to be conscious of the external ones. However, as Husserl then continues, *such a theory is completely nonsensical*. (Zahavi 2018: 56-57, my emphasis)

According to Husserl, the theory makes no sense because the idea that consciousness is some kind of box containing representations that resemble external objects completely ignores the problem of how we are "supposed to know that the (mis)representations are in fact (mis)representations of external objects" (Zahavi 2018: 57) The following excerpt from Husserl's lecture leaves little doubt as to his anti-representationalist position:

> The ego is not a tiny man in a box that looks at the pictures and then occasionally leaves his box in order to compare the external objects with the internal ones etc. For such a picture observing ego, the picture would itself be something external;

> it would require its own matching internal picture, and so on *ad infinitum* (Husserl 2003: 106).

The above citation becomes even more pertinent when we view it in light of Kosslyn's influential 'two-storey story', which not only postulates internal representations at both the level of description *and* the underlying explanatory level; the explanatory representational posits appear to be precisely the picture-like representations Husserl was already arguing against more than a century ago. Yet, even if we ignore Husserl's grounds for dismissing the idea of internal picture-like representations, Kosslyn's account remains problematic for a number of different reasons still.

3.11 Additional problems with 'two-storey stories'

As said, Kosslyn is one of the main protagonists in the imagery debate. He defends the claim that the format of the internal representations underlying visual imagery and imagination are image- or picture-like. His empirical research wants to provide evidence for this claim. Yet, this approach skips over the following two questions: first, what scientific reasons do we have for believing that our experiences of visual imagery and imagination (as well as other forms of cognition) are 'underlain' by internal representations at all, regardless of their format? And second: assuming there are such 'underlying' representations, how, exactly, are they supposed to be doing their explanatory work?

As to the first question, we find no answer in Kosslyn's work, nor anywhere else. Cognitive science literature is rife with representation-talk, but I am as yet still to come across an account in which it is convincingly argued that the belief in the existence of internal representations is the result of scientific research, rather than a base assumption. As François Tonneau observes in this regard:

> In cognitive psychology…representational attributions are not the result of, but the prerequisite for, theoretical development. Representations are invoked even *before* the theory starts.(Tonneau 2011/2012: 338)

The question imposes itself as to what it is about the idea of internal representations that makes it so readily acceptable as a fundamental assumption for cognitive scientists studying forms of off-line cognition. I want to suggest a possible answer, which at once addresses the question of how internal representations are thought to be doing any explanatory work.

3.12 Explaining the representational appeal: 'like causes like'

As already indicated above, I believe much of the internal-representation-idea's appeal derives from a conflation between the descriptive and the explanatory level. It cannot be a mere coincidence that theorists like Kosslyn postulate sub-personal image-like representations to explain subjective image-like representations (visual imagery). It is plausible that this conflation is best understood as an expression of the idea that, for an explanans to qualify as explanatorily satisfying, the explanandum must somehow still be sufficiently recognizable in the explanans. Here, I want to make the – admittedly speculative – suggestion that there is something at work here which is, within anthropology, referred to as the 'law of similarity'. Since long time, it has been observed by ethnologists that human pre-scientific thought is governed by what are called 'laws of sympathetic magic' (Rozin & Nemeroff 2002: 201)[71]. Usually, three such laws are being distinguished: the 'law of contagion', the 'law of similarity' and the 'law of opposites'. According to the second of these laws, human thought is prone to adhere to the principle that 'causes resemble their effects', or that 'like causes like'. To be clear, these aren't just heuristic principles adhered to by people living in traditional societies, far removed from modern academies, experiment rooms and laboratories. These principles are just as well exercising their influence on the well-educated contemporary Westerner. Rozin & Nemeroff (2002) cite as an example the practice of homeopathy: "This cause-effect likeness principle is at the foundation of the tradition of homeopathic medicine"(Rozin & Nemeroff 2002: 204). Another example these authors mention is the widespread intuition

[71] Rozin and Nemeroff refer in particular to the work of pioneering ethnologists Edwin Tylor, James Frazer and Marcel Mauss.

that, because a certain disease is highly resistant to treatment (e.g., AIDS), the underlying agent causing the disease (e.g., HIV virus) must have the same potent and indestructible properties (which, in the case of HIV, happens to be untrue). My – again, tentative –suggestion, then, is that this principle is to some extent also at work at the foundation of mainstream representationalist cognitive science, in which representational phenomena are assumed to be somehow causally explainable by representational entities.

In the particular case of the explanation of representational cognition like visual imagery and imagination, this principle manifests itself in the assumption that visual depictive representation (imagery) requires depictive representation in its explanation[72]. And it is with regard to the question as to *how* these underlying representational entities are supposed to be doing their explanatory work that reification comes in. After all, assuming internal sub-personal representation to explain experienced representation (the "tip of the iceberg") is in itself not enough to provide the sort of causal explanations that science demands. Hence the postulation of internal *object-like* entities which are at once causally efficacious (in virtue of being physically implemented), as well as sufficiently similar to the explanandum, i.e., representational. Yet, in spite of those theorists that are attracted to such a view, it raises more problems than it solves. Not only do we not know how to naturalistically account for the idea of internal content-carrying representations, even if we would have such an account, we would still be left with the question as to how anything representational (internal or not) can causally explain anything representational. How can sub-personal entities

[72] My critical interpretation of this explanatory scheme is in some respects similar to what Pessoa, Thompson & Noë have called 'analytical isomorphism': "Analytical isomorphism is the idea that successful explanation
requires there be an isomorphism (one-to-one correspondence) between the phenomenal content of subjective experience and the structure or format of the underlying neural representations. This idea involves conflating properties of what is *represented* (representational contents) with properties of the *representings* (representational *vehicles*)." (Thompson 2007: 272; see also Pessoa, Thompson & Noë 1998). However, my claim wants to be more general and does not hinge on the idea of isomorphism, nor on the notion of phenomenal content. On my interpretation, the similarity is more general and is supposed to apply more widely in that it pertains to the alleged representational nature itself of both the explanandum and the explanans. So my interpretation is supposed to not only apply to Kosslyn's picture-like format of the underlying internal representations, but equally to Pylyshyn's propositional format. Both assume in their own way a representational explanatory base for representational explananda, in this case, mental imagery.

that somehow say or depict how things stand (or, in case of imagination, *could* stand), explain our subjective and active conscious experience of, e.g., imagining how things could be? When somebody asks me to imagine the Eiffel Tower, it is *I* who am doing the imagining, and it is *I* who knows how to do it such that I don't imagine the Colosseum (even though I do not know how I do it). How does the postulation of sub-personal, object-like internal representations help to explain this capacity? And perhaps most pressingly: what does it mean, exactly, to say of *any* object, internal or external, that it *represents*? Shouldn't we be clear on this first, before we start to raise questions as to how to naturalize representation? These are fundamental questions that within mainstream cognitive science remain not so much unanswered, as simply unasked.

Before returning to our main topic, namely the question of whether non-representationalist approaches to cognition are principally excluded from off-line cognition research, I want to briefly summarize the previous section.

3.13 Intermediate summary

I have distinguished three reasons for rejecting the idea that Husserl – and other phenomenologists in his wake – can be seen as a forerunner for the cognitive revolution and its representation-invoking explanatory framework, which has become standard within cognitive science. For although we encounter *some* notion of mental representation in Husserl's phenomenological inquiries, this notion differs crucially from the one we find in mainstream cognitive science in that, first, it is *descriptive* and not explanatory, second, it is *subjective* and *experiential* instead of objective and sub-personal, third, it refers to an *activity* and not to something object-like. As we've also seen, for Husserl, the notion of representation first of all pertains to conscious activities that relate to the absent (remembering, imagining, viewing pictures) which are all understood as modifications of perception (Marbach 1993: 1). This latter point is important, for it suggests the following: If non-representationalist approaches might be preferred over representationalist approaches for the study of on-line

cognition(including perception), and if, at the same time, 'representation-hungry' off-line cognition can be understood as a modified form of on-line cognition (perception), then perhaps non-representationalism has something important to say about these modifications of on-line cognition as well. Otherwise put: if off-line cognition is to be understood on the basis of on-line cognition, and if on-line cognition does not require invoking internal representations, then perhaps we can do without internal representations when it comes to off-line cognition as well. This line of thought will be further pursued below.

3.14 Turning representation-hungry problems on their heads

Interestingly, it is Andy Clark himself who, in later work, offers a tentative exploration of precisely this idea, namely that forms of off-line cognition are much more indebted to on-line perceptual engagements than is usually assumed. What is more, according to Clark, it is precisely the on-line engagement with *public* representations that allows for off-line cognition to get up and running. Compared to Clark and Toribio 1994, Clark 2005 finds the whole idea of excluding representation-hungry cognition from non-representationalist approaches "less compelling than I once believed." (Clark 2005: 233) Furthermore, Clark argues that the distinction between on-line and off-line cognition is much less rigid than is often assumed. First, he underwrites the idea that

> in just about all cases, we find elements of each and a constant seamless integration of the two. The need to integrate real-time action with ongoing planning and reason is itself, it is argued, a reason to prefer a unified dynamical approach that treats perception, reason, and action in essentially the same terms. (Clark 2005: 234)

Second, Clark emphasizes that

> even imagination-based, problem-situation-decoupled reason is fully continuous with the other cases, because our imaginative routines are themselves body-

> based, exploit egocentric coordinate spaces, (re)deploy the same perceptuo-action-oriented inner states, and so on. (Clark 2005: 235)

Indeed, neuroscientific research has gathered evidence that there are, neurologically speaking, close similarities between current acts of imagining or remembering and previous perceptual activity. Various experiments found "that imagined patterns and seen patterns produced similar waveforms, supporting evidence for the claim that the visual cortex is activated in a similar manner during both imagination and perception." (Vitrano 2012: 2)[73] This result also supports the proposal of understanding imaginings as "perceptual reenactments" (Hutto 2008: 79; see also Currie 1995; Currie and Ravenscroft 2003; Prinz 2002). Rather than causally relying on some content carrying sub-personal representations, on the reenactment account, visually imagining something is understood as a kind of enacted "virtual perception" (Degenaar & Myin 2014: 3644). As Evan Thompson puts it:

> [V]isualizing is not an experience in which we seem to see or have a mental picture. Visualizing is rather the activity of mentally representing an object or scene by way of mentally enacting or entertaining a possible perceptual experience of that object or scene. (Thompson 2007: 279)[74]

And – so it is argued – since perception does not require invoking internal representations, neither does its reenacted counterpart, i.e., imagery. Of course, across the board representationalists will object that the neural similarities between perceiving and imagining can be taken as supporting precisely the opposite claim that not only off-line cognition like imagining involves representations, but on-line perception as well. Indeed, as Degenaar & Myin point out,

> [i]f one is a representationalist about perception, imagery, understood as virtual perception, will obviously involve representation too. If, however, one combines a

[73] See, for instance, Gelbard-Sagiv et al., 2008.

[74] In their recent book, Di Paolo, Buhrmann & Barandiaran write: "Instead of activating some internal image-like neural pattern stored in your brain and then "looking at it," you are closer to acting or, as in a theatre play, *enacting* the visual experience (…). It is an act of presenting something to yourself again, or re-presenting." (Di Paolo, Buhrmann & Barandiaran 2017: 28.)

> virtual perception take on mental imagery with a view of perception as non-representational, one will be led to a non-representational view of imagery.... Crucially, if an account of perception in the presence of a stimulus is non-representational, the account of imagery as a kind of perception in the absence of the stimulus, may be so too. (Degenaar & Myin 2014: 3644-5)

But although the neuroscientific findings may be interpreted either way, non-representationalists seem to have at least this much in their favor that theirs is the more parsimonious account. If a good representation-eschewing case can be made for forms of on-line cognition like perception, it is up to the representationalist to argue why internal representations are nevertheless explanatorily indispensable.

In addition to these two reasons[75] for relativizing the on-line/off-line distinction, however, according to Clark, there is still another, "more fundamental" (Clark 2005: 235) argument for assuming a close affinity between off-line cognition and on-line perceptuo-motor engagements. The argument is based on considerations about the role of public representations, and especially language, for off-line forms of cognition. In short, it is our on-line engagements with public representations like words and sentences (but also diagrams, sketches, models, and, in general, all objects that serve as a surrogate for some actual or potential

[75] It should be noted that, in addition to these "two standard replies" (Clark 2005: 234), recently, the on-line/off-line cognition distinction has been reassessed from both a Heideggerian as well as a Gibsonian perspective. Michael Wheeler relates the distinction to Heidegger's 'ready-to-hand/present-at-hand' distinction. Ludger van Dijk and Rob Withagen (2016) follow Wheeler here, but further interpret the on-line/off-line distinction as inextricably linked to a certain time conception (time as abstract time) which is said to pervade mainstream psychology (see, respectively, Wheeler 2005 and van Dijk & Withagen 2016). However, although these readings have their merits, they are each in their own way overelaborated with regard to the on-line/off-line cognition distinction and, therefore, prone to give a somewhat distorted picture of the distinction as it originally occurs in the work of Clark and others. van Dijk & Withagen's focus on temporality overlooks the fact that forms of off-line cognition are not only defined by temporal absence, but also spatial absence. The distinction can't therefore be "dissolved" (van Dijk & Withagen 2016: 7) by considerations of underlying time conceptions alone. Furthermore, it is doubtful that Clark & Toribio would agree with van Dijk & Withagen's claim that 'object-permanence' constitutes "the basic representation-hungry case" (van Dijk & Withagen 2016: 6). Wheeler, on the other hand, appears to be drawing a simply too far-fetched parallel between Heidegger's existential-phenomenological concepts of the ready-to-hand and the present-at-hand. Not only is it far from clear how these Heideggerian notions can be taken to pick out different forms of *cognition*, they also seem to be notions that both pertain to Dasein's dealings with what is *present,* rather than absent. This seems obvious from the fact that these notions are supposed to relate to two different ways in which the world, and in particular, a tool, can *present* itself to Dasein.

real-world state of affairs) that lie at the basis of off-line forms of cognition, such as reasoning and planning. According to Clark, the relation between on-line engagements with representations and the development of our more decoupled cognitive capacities is to be understood historically in terms of a "gradual coevolution" (Clark 2005: 238). To illustrate the role of public language symbols for off-line cognition, Clark cites an experiment by Hermer and Spelke, in which it is shown how language competent participants are much more successful at remembering an object's location than rats or prelinguistic infants. Clark explains these results as follows:

> Experience with public language symbols [...] allows us to direct and distribute attention in new ways. And it does so by in effect creating a special kind of surrogate situation: one in which what is otherwise unavailable is not the visual scene itself, but a particular way of parsing the scene into salient components and events. (Clark 2005: 239)

With regard to the experiment above, because linguistically competent participants have access to surrogates (words) that stand for certain environmental aspects, they are more successful at remembering the location of an object because they are capable of linguistically representing cues that would escape the attention of the illiterate animal or infant. For example, because the participant can make use of color-words, she can describe something as "to the right of the green wall", thus combining spatial information (which is presumed to be the exclusive kind of information rats rely on) with a different kind of information – color-information – which apparently escapes the attention of rats and prelinguistic infants. Clark further discusses Peter Carruthers' interpretation of these sorts of experiments, which are taken to involve the combination of different kinds of information:

> Perceptually encountered or recalled symbols and sentences act, according to Carruthers, like inner data structures, replete with slots and apt for genuine inner combinatoric action. This combinatoric action allows information from otherwise encapsulated modules to enter into a unified inner representation. (Clark 2005: 239)

In sharp contrast with Clark's earlier defense of internal representations (Clark & Toribio: 1994), here, Clark rejects Carruthers' idea of translating the experiment in a representationalist vocabulary. In a rather dramatic reversal, Clark 2005 writes:

> Contra Carruthers, then, I think we may conceive *perceptually encountered* or recalled symbols and sentences as acting less like inner data structures, replete with slots and apt for genuine inner combinatoric action, and more like cheap ways of adding task-simplifying and attention-reconfiguring structure to the *perceptual scene*. (Clark 2005: 240; my emphases)

The phrases "perceptually encountered" and "perceptual scene" deserve emphasis, as they affirm the primacy of perception over other forms of cognition, and in particular those forms of cognition that Clark had earlier deemed 'genuine'. Ironically, Clark fails to notice that, on his newer account, these forms can still be properly labelled 'representation-hungry', provided, of course, that 'representation' now refers to ordinary public/external representations, and no longer to the hypothetical private/internal entities (which, we should repeat, are modelled on the external ones anyway).

3.15 Towards a dynamical account of on-line and off-line cognition

Clark is, of course, not the only one to hold that the presence of linguistic symbols, as well as other public representations, have a crucial role to play in the development of our higher forms of cognition[76]. The idea is nowadays endorsed by many theorists with sympathies towards E-approaches to cognition and it also forms, for instance, a central theme in the work of radical enactivist authors like Hutto & Myin. Viewing forms of off-line cognition in a developmentally close relation to our (external) representation-using practices opens undoubtedly promising perspectives. It should be noted, however, that it also raises some questions of its own. To see this, I want to refer to a passage from the already

[76] For other accounts stressing the role of external representations for off-line cognition, see for instance Kirsch 2010.

mentioned phenomenological work *Mental Representation and Consciousness*, by Edouard Marbach. As already indicated, Marbach's phenomenological work focusses on the various ways in which consciousness engages in acts of re-presenting, understood in Husserl's sense as modifications of acts of perceiving (or presenting). Fully aware of the routine-like fashion in which contemporary cognitive science invokes internal representations, especially when it comes to cognition involving the absent (i.e., off-line cognition), he writes:

> Reference to something in its absence is philosophically not to be explained by having recourse to a putative vehicle of representation of one or another "format", to some "internally present representative" of the absent object (the past event, the merely imagined scene, the anticipated meeting etc.). Instead, the way of reference must philosophically be analyzed *in the first place* by means of uncovering the novel intentional complexities pertaining to the mental activities themselves which, in virtue of being performed, refer to something absent over and against something in its present givenness. ...[S]uch referring may occur with or without a public vehicle of representation, and it may occur with or without a vehicle of representation that is itself purely mentally represented. (Marbach 1993: 10)

The non-representationalist attitude we find in the first sentence is clearly reminiscent of Husserl's dismissal of internal representations, as we've found it in Zahavi 2018. But it is what comes after this sentence that is especially of interest here. Marbach argues that, although "mental activities referring to the absent" (off-line cognition) may involve both external as well as internal representational vehicles, these mental activities are first of all in need of an analysis in their own right, that is, *qua mental activity*. As I understand it, the point is that, before we can even begin to try to *explain* off-line cognition in terms of representational objects (whether they are external or internal), we should first try to better understand the "intentional complexities pertaining to the mental activities themselves" because it is only through these activities that an object can acquire the status of 'representational object'. What this means, very simply, is that *all* representational objects depend for their representational status on certain cognitive activities, or perhaps better: cognitive capacities.

Viewed in this light, the dramatic reversal pertaining to Clark and Toribio's notion of representation-hunger not only concerns the 'location' of the representations (internal vs. external), in addition, the dependency relation between off-line cognition and representational vehicles is on Marbach's account also reversed. Here, a public object can only acquire the status of a representation, precisely *in virtue of certain off-line cognitive capacities*, and in particular our imaginative capacities, as well as our capacity to abstract away from irrelevant details. Otherwise put, for something to be standing for something absent or abstract, it must be doing so for someone. But it can only do so if that someone has already the capacity to relate to it as something it is not, which requires precisely the kind of off-line capacities Clark and others want to explain by appealing to these public representational objects, and especially the capacity of imagination. This would mean that forms of off-line cognition like imagination aren't representation-hungry at all, but that precisely the contrary is the case, namely that *representations are imagination-hungry*. *Prima facie*, this leaves us with a circularity: on the one hand, it is claimed that our off-line capacities depend on our public representation-using practices. On the other hand, these practices are themselves dependent on our off-line cognitive abilities.

3.16 Vicious and virtuous circularities

Rather than viewing accounts that stress the importance of public representations for off-line cognition as viciously circular, my suggestion would be to regard these accounts, not as themselves circular, but rather as pointing towards a real non-viciously circular system. Instead of reducing the issue to an all-or-nothing 'chicken or egg' affair, a much better, as well as more plausible idea would be to regard the development of forms of off-line cognition and the development of public representation use as both co-dependent and mutually reinforcing. On-line engagements with certain objects require some basic form of imagination in order for these objects to be usable as standing for something absent. But once these objects have acquired their representational status –

something which can only be established within a proper socio-normative context – further engagement with these objects stimulates imaginative capacities, which in turn facilitates the further use and (re)production of public representations, and so on. In terms of on-line vs. off-line cognition, we might say that our on-line perceptuo-motor engagements with public representations, and especially language, stimulate forms of off-line cognition (e.g., imagining the absent, remembering the absent, reasoning about the absent), and that, in turn, this has an effect on our on-line perceptuo-motor engagements as well: we come to see the world quite literally in different terms, and we accordingly act differently upon it. In the final analysis, then, rather than an ontological distinction, on-line and off-line cognition is best understood in terms of a coupled system in which relations of co-dependency and mutual reinforcement (positive feedback) provide the system with its proper dynamics. Indeed, this gives us precisely no reason to think that a non-representational dynamical systems approach is bound to fall short here. True, the full ontogenetic story will still involve representation-talk, but it will not involve the hypothesized intracranial content-carrying entities that are by now overpopulating mainstream cognitive science. And contrary to the cognitivist's internal representations, on the developmental account, here we have representations we can actually *see*.

3.17 Concluding Remarks

The idea that higher, off-line forms of cognition necessarily require internal-representation-invoking explanations, and that, therefore, non-representationalist approaches are by default incapable of dealing with these forms of cognition, is a widespread assumption of mainstream cognitive science. As we've seen, this assumption can't only be found in classic cognitivist approaches – where *all* forms of cognition are thought to be representational anyway – but in (more conservative) E-approaches as well. Furthermore, to the extent that only these higher, off-line forms of cognition are assumed to be genuinely cognitive in nature, the non-representationalist can't even be said to

investigate real cognition to begin with. As we've also seen, despite existing efforts, we have as yet no good argument for accepting these assumptions. On closer examination, these assumptions seem to derive, not so much from empirical or scientific considerations, but from less estimable sources. I've tried to show that the idea that off-line cognition must be explained in terms of internal representations is built on a conflation of the level of description with the level of explanation. Furthermore, I have tentatively suggested that an explanation for this conflation itself should be sought in our cognitive predisposition to assume that causes must resemble their effects, a principle referred to by ethnologists as the 'law of similarity'. In any case, however, denying that forms of off-line cognition like memory and imagination can only be explained in terms of operations over internal content-carrying vehicles does not require denying that these forms of cognition can be *described* representationally, i.e., as capacities to re-present something which is in some sense absent. Husserl's phenomenological descriptions of these re-presentational capacities can therefore still be of great value to the non-representationalist, especially since – despite what Dreyfuss once believed – Husserl himself was a staunch opponent of the sort of representational entities that mainstream cognitive science puts forward as explanatory posits. Representation, both for the Husserlian phenomenologist, as well as the non-representationalist cognitive science theorist, should not be smuggled over from the side of the explanandum to that of the explanans, but should first of all be approached as an interesting research subject in its own right. For rather than explaining our off-line capacities, it may turn out that representation is itself in need of an explanation that relies on these capacities (in particular, imaginative capacities) already being in place, at least in their latent form. To be sure, it is very likely that representations will have an indispensable role to play in our best theories of higher forms of cognition such as memory and imagination, provided that 'representation' here refers first of all to socially shared external entities, especially linguistic entities. And, as explained in the above, these theories will need to take into account a certain loopy dynamics of mutual dependence, which may aptly be referred to as a virtuous circularity. However, as long as cognitive scientists keep assuming that these

external representational entities must themselves be causally explained by internal entities that resemble the external ones in essential respects (e.g., by postulating a Language of Thought), the circularity they are faced with can only be vicious.

References

Adams, F. (2010a). Embodied cognition. *Phenomenology and the Cognitive Sciences: Special Issue on 4e Cognition, 9*, 619–628.

Adams, F. (2010b). Why we still need a mark of the cognitive. *Cognitive Systems Research 11,* 324–331

Adams, F., and Aizawa, K. (2008). The Bounds of Cognition. Malden, Mass.: Blackwell.

Adams, F., and Garrison, R. (2013). The mark of the cognitive. *Minds & Machines, 23*, 339–352.

Beer, R. (1995). A dynamical systems perspective on agent-environment interactions.

Artificial Intelligence, *72*(1-2), 173–215.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J. A. Meyer, J. Pollack, & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior,* 421–429. Cambridge, MA: MIT Press.

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

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

Clark, A. (1996). *Being There*. Cambridge, Mass.: MIT Press.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351.

Clark, A. (2005). Beyond the flesh: Some lessons from a mole cricket. *Artificial Life, 11*(1-2), 233–244.

Clark, A., and Toribio, J. (1994). Doing without representing? *Synthese, 101*(3), 401–431.

Currie, G. (1995). Visual imagery as the simulation of vision. *Mind and Language 10*(1–2), 25–44.

Currie, G. and Ravenscroft, I. (2003). *Recreative Minds*. Oxford: Oxford University Press.

Degenaar, J., and Myin, E. (2014). Representation-hunger reconsidered. *Synthese, 191*(15), 3639–3648.

Dennett, D. C. (1978). *Brainstorms.* Cambridge, MA: MIT Press.

Di Paolo, E. A., Buhrmann, T., and Barandiaran, X.E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford University Press.

Dreyfus, H. L. (ed.) (1982). *Husserl, Intentionality, and Cognitive Science.* MIT Press/Bradford Books, Cambridge.

Edelman, S. (2003). But will it scale up? Not without representations. *Adaptive Behavior, 11*(4), 273–275.

Epstein, R. (2016). The empty mind. Published online by Aeon. URL= https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Gelbard-Sagiv, H. et al. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, *322*(5898), 96–101.

Husserl, E. (2003). *Transzendentaler Idealismus. Texte aus dem Nachlass (1908–1921).* Husserliana 36. R. Rollinger (Ed.). Dordrecht: Kluwer Academic Publishers.

Hutto, D. D. 2008. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Hutto, D.D., and Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Hutto, D. D., and Myin, E. (2017). *Evolving enactivism: Basic minds meet content.* Cambridge, MA: MIT Press.

Hutto, D.D., Myin, E., Peeters, A., and Zahnoun, F. (forthcoming). The cognitive basis of computation: putting computation in its place. Forthcoming in Sprevak, M. and Colombo, M. (Eds.), *The Routledge Handbook of the Computational Mind*. Routledge.

Kirsh, D. (2010). Thinking with external representations. AI & Soc (2010) 25: 441. https://doi.org/10.1007/s00146-010-0272-8

Kosslyn, St. M. (1978). Imagery and Internal Representation. In Rosch, E. and Lloyd, B.B. (Eds.), *Cognition and Categorization.* Hillsdale: Lawrence Erlbaum Associates, 217-257.

Kosslyn, St. M. (1980). *Image and Mind.* Cambridge, Mass.: Harvard University Press.

Kosslyn, St. M., Thompson, W. L., and Ganis, G. (2010). *The Case for Mental Imagery.* Oxford University Press.

Marbach, E. (1993). *Mental Representation and Consciousness: Towards a Phenomenological Theory of Representation and Reference*. Dordrecht: Kluwer Academic Publishers.

McIntyre, R. (1986). Husserl and the Representational Theory of Mind. *Topoi* 5, 101-113.

Pessoa, L., Thompson, E., and Noë, A. (1998). Finding out about filling-in: a guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences 21*, 723-802.

Pitt, D. (2017). Mental Representation. *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/spr2017/entries/mental-representation/.

Preston, B. (1994). Husserl's non-representational theory of mind. *The Southern Journal of Philosophy, 32*(2), 209-232.

Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin, 80*(1), 1-24.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Ramsey, W. M. (2017). Must cognition be representational? *Synthese*, *194*(11), 4197-4214.

Rowlands, M. (2010). The new science of the mind. Cambridge, MA: MIT Press.

Rozin, P. & Nemeroff, C. (2002). Sympathetic magical thinking: the contagion and similarity "heuristics". In T. Gilovich, D. Griffin & D. Kahnemann (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Shapiro, L. (2014). Review of *Radicalizing Enactivism*. *Mind, 123*(489), 213–220.

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

Tonneau, F. J. (2011/2012). Metaphor and truth : a review of representation reconsidered by W. M. Ramsey. *Behavior and Philosophy (Online) 39/40,* 331-343.

van Dijk, L. and Withagen, R. (2016). Temporalizing agency: Moving beyond on- and offline cognition. *Theory & Psychology, 26*(1), 5–26.

van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy,91*(7), 345–381.

van Rooij, I., Bongers, R., and Haselager, W. (2002). A non-representational approach to imagined action. *Cognitive Science, 26*, 345–375.

Vitrano, D. M. (2012). Comparing perception and imagination at the visual cortex. Dickinson College Honors Theses.

Wheeler, M. (2005). *Reconstructing the cognitive world*. Cambridge, MA: MIT Press.

Wilson, R. A., & Foglia, L. (2011). Embodied cognition. *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. (ed.), URL = http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/.

Zahavi D. (2018). Brain, Mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism. *Husserl Studies, 34*(1), 47-61.

Zarkadakis, G. (2016). *In our own image.* UK: Rider Books.

4 Multiple Realization: A Thesis with Identity Issues

Abstract

It is commonly held that the multiple realizability of the mental rules out a potential strict identity relation between the physical and the psychological. In recent years, important new work has been done on the subject of the relation between multiple realization and identity theory. Nevertheless, what remains overlooked by these newer accounts is that the argument from multiple realization against identity is founded on an ambiguity. At the heart of the multiple realization argument lies an entanglement of two very different notions of identity, notions which need to be understood in light of two distinct ways of identifying objects. As I will hope to show, properly evaluating the argument from multiple realization against strict identity requires first of all untangling these two notions of identity. The disentanglement leaves the argument from multiple realization facing a dilemma: either be a deductively valid argument, but give up on empirical aspirations; or be an empirically substantiated argument, but accept compatibility with a strict identity thesis.

4

Multiple Realization: A Thesis with Identity Issues

4.1 Introduction: Strict Identity 2.0

In the early sixties, Hilary Putnam introduced the idea that the same mental state kinds are realizable by different physical kinds. Ever since, this idea – which came to be known as 'multiple realization' – has been widely accepted as both an argument against, as well as a superior alternative to the classic identity theories as advanced by Ullin Place (1956), Herbert Feigl (1958) and John J.C. Smart (1959). As William Bechtel and Jennifer Mundale put it: "MR[77] has become orthodoxy in the philosophy of mind" (Bechtel & Mundale 1999: 176; see also Bickle 2003: 131).

At the same time, however, we see that identity theory has never really been laid to rest. In recent years, proposals of assuming a strict mind-body identity relation (i.e., that mind and body are ultimately one and the same) have resurfaced within philosophy of mind, but also within philosophy of cognitive science, where it figures centre stage in current E-accounts of cognition. Sensorimotor theories of consciousness, for instance, postulate a strict identity between sensorimotor activity and phenomenal experience. Alva Noë holds that "perceptual experience *just is* a mode of skilful exploration of the world." (Noë 2004: 194) And in addressing the Hard Problem of Consciousness, proponents of so-called Radical Enactive Cognition (REC for short) suggest adopting a strict identity thesis as perhaps the *only* viable approach to the Hard Problem. Michael Kirchhoff and Daniel Hutto, for instance, claim that, because of an assumed strict identity relation between the phenomenal and the physical, explaining why correlations hold isn't so much hard as impossible: "there is simply no problem to solve here"

[77] MR: multiple realization. I'll use this abbreviation throughout the paper.

(Kirchhoff & Hutto 2016: 308)[78]. Indeed, as the classic identity theorists already emphasized, if the phenomenal-physical relata are strictly identical, there are no correlations between them: "You cannot correlate something with itself." (Smart 1959: 142; see also Feigl 1958: 70)

To be sure, although the strict identity notion at work in these newer theories is the same one that has been advanced by the early brain state identity theorists, enactive proposals about the relation between the phenomenal and the physical are at the same time also very far removed from the classic brain-centric identity theories. The classic idea that mental states can (only) be identified with brain states is traded in for the very different suggestion that, when it comes to the relata of the identity relation, we should, as Daniel Hutto and Erik Myin say, "go wide"[79]. Conscious experience is here no longer identified with internal or neural processes, but instead with bodily (including neural) processes in spatiotemporally extended interactions with environments. In this vein, Tim Ingold stresses that "body and mind are not… two separate things but two ways of describing the same thing – or better, the same process, namely the activity of the organism-person in his or her environment." (Ingold 2001: 240)

Identity claims such as these will strike many as non-starters. After all, Putnam provided us with a strong argument against the mind-brain identity theorist already half a century ago: The same type of mental state can be realized by different physical state types, so the mental can't be identified with the physical. Do we have reasons to think that these more recent wide identity proposals will hold up better against MR? In the following, I want to reconsider the so-called argument from MR against identity and show that it lacks the conceptual means to form a threat to a properly understood identity theory, regardless of how wide we take the physical states to be. Indeed, my claim will be that MR should never have been accepted as a defeater of a potential identity relation between the mental and the physical.

[78] See also Papineau (1998): "I think physicalism is best conceived as a thesis of identity between conscious properties and material properties, and identities need no explaining." (Papineau 1998: 373)
[79] See Hutto & Myin 2013: 151.

4.2 MR: Identity, similarity and difference

In recent years, important new work has already been done on this subject. In the following, I will take a closer look at some of these more recent accounts. However, as I will try to make clear, what remains overlooked is that the MR argument against identity draws on an entanglement of two very different notions of identity. Properly evaluating the argument from MR against strict identity requires a disentanglement of these two notions. I'll begin my argument, however, with some preliminary recapitulations.

Although Shapiro (2000) rightly notes that there is no one accepted interpretation of what the idea of MR is supposed to hold exactly, on a common interpretation (going back to Putnam), MR is the idea that one and the same type of mental entity[80] can be realized by different types of physical entities. This being said, it should already be emphasized that MR is not a thesis about the mind-brain relation *per se*: to the extent that MR is also taken to apply to non-mental functional entities – as to my knowledge is done by all authors – this formulation is too restrictive. On its broadest functionalist construal, MR can be formulated as the idea that one and the same type of functional entity can be realized by different types of physical entities. And just as MR is true of things like mousetraps, carburetors and corkscrews, it is also held to be true of mental entities like pain sensations. Of course, for the functionalist, the mental simply is a functional category.

At the same time, this is supposed to entail a deductive argument against the brain-state identity theorist. Recall this oft-quoted passage by Putnam:

> Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian

[80] I deliberately use the umbrella term 'entity' here so as to comprise states, processes, structures or properties. Depending on whose definition one considers, MR is predicated of either one of these ontological categories. My use of the term 'entity', then, is supposed to encompass them all.

> brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. (Putnam 1975: 436)

Clearly, identity theory is here understood as claiming that a single mental type (pain) is identifiable with a single physical kind. On this construal, it follows that, if MR is true, identity theory is false. This standard version of the MR argument against identity is schematized by Polger as follows:

> Ψ stands in R to physical state P_1[81] [in creature C_1.]
>
> Ψ stands in R to physical state P_2 [in creature C_2.]
>
> Therefore, Ψ is not identical to any single physical state type. (Polger 2013: 865)

This, at best, has the appearance of a deductively valid argument. It is, however, question-begging in that it already assumes that Ψ is in fact realized by a physical state. It also assumes that the relation of 'being identical to' is supposed to hold between types. But perhaps this is not the best way to interpret the claims of an identity theorist. Furthermore, the argument has a built-in redundancy: the assumed multiplicity of the realizations actually has no role in the argument. The realization relation itself is enough to rule out identity. If the relation between a mental state type and a physical state type is that of realization, then what difference does it make that the same mental state type can be *multiply* realized? Polger, who also emphasizes this point, therefore maintains that "the argument…would better be called the *realization argument*, for the fact of multiplicity of the realizations is superfluous." (Polger 2013: 866) As we'll see, Polger himself goes to great lengths to defend a different reading of MR which safeguards it from the charge of begging the question, and which affirms its status as a genuine empirical thesis. What I want to argue is that, orthogonal to the above criticisms of MR, the deductive MR argument against identity is inherently confused with regard to the notion of identity. As already indicated above, properly evaluating the MR argument against identity, as well as the MR thesis itself, requires disentangling two notions of identity that are continuously

[81] 'Ψ' standing for the mental entity type, 'R' for the realization relation and P for the physical entity type. See Polger 2013: 865.

being mixed up in the literature on MR, and in the way the thesis' relation to identity theory is being conceived.

To see more clearly what I mean when I say that there is a confusion of two notions of identity at work here, a first requirement is to reveal and distinguish these two notions. I'll do so in the next section.

4.3 Two modes of identification: P-Identification & C-Identification

At the beginning of *Sameness and Substance Renewed*, David Wiggins writes:

> Let the philosopher elucidate *same*, *identical*, *substance*, *change*, *persist*, etc., directly and from within the same practices as those that an ordinary untheoretical human being is initiated into. (Wiggins 2001: 2)

I will take Wiggins' incitement to heart by relating the two tangled up notions of sameness to two different practices of identification that I believe are underlying these notions. Provisionally, we could refer to these two different notions of identity as strict identity or one-and-the-sameness on the one hand, and relative or categorical identity on the other. As an illustration of the first kind of identity, we could think of the sense in which we attribute sameness in an assertion like "The knife I am holding is the murder weapon." The knife and the murder weapon are one and the same thing. An example of the second kind of identity can be found in the way we attribute identity to an object in assertions like "The thing you are holding is a knife." Here, identity refers to *what* something is, which is always a matter of relating it to its proper category (hence relative or categorical identity). As said, however, in compliance with Wiggins, I want to connect these two different notions of identity to our different identificatory practices. My reason for doing so is twofold. On the one hand, an examination of our identificatory practices leads to a more explicit and demarcated conception of the two different identity notions. On the other hand, and more importantly, relating these notions of identity back to their respective practices is also meant to resist, what Wiggins calls, the "realist myth of the *self-differentiating object*

(the object which announces itself as the very object it is to any mind, however passive or of whatever orientation)" (Wiggins 2001: 150-151). What we need to keep in mind is that all notions of identity are also relational in the sense that they involve a relation to an identifying subject. However, since acts of identification are usually unproblematic, and therefore stay unobserved *qua* acts, we tend to lose sight of the fundamental fact that identification is never simply the passive registration of identities out there, but a structuring activity in which *we* bring something to bear on the world.

The important point, then, is that humans identify in two very different ways, which gives rise to two very different notions of identity or sameness. Below, I will further analyse our provisional distinction between strict and relative identity by analysing our distinct ways of identifying.

4.3.1 P-Identification and P-Identity

First[82], we identify in the sense of determining sameness as 'being one and the same'; that is, we identify something *with*, as opposed to identifying something *as*. These acts of identification usually happen pre-reflectively. When I see a car parked in front of the house, I do not need to perform some special intellectual act to see it as the same car as the one I parked there the night before. I just see *my* car. Determining sameness in this sense is usually unproblematic, although, on occasion, difficulties might arise. This is inevitable because of the perspectival nature of our acquaintance with objects, both in space and time. First, because it is usually impossible to uninterruptedly observe an object, it might become problematic to determine sameness of an object after we've temporarily lost track of it. Doubts might arise as to whether some object at t_1 is one and the same at t_2, for instance because it has changed beyond recognition (the caterpillar at t_1 and the butterfly at t_2), or precisely because it still looks the same at t_2 as it did at t_1,

[82] The order in which I treat both modes of identification is completely arbitrary and certainly not supposed to reflect a historical order, or any other order for that matter.

although we expected it to have changed (someone you haven't seen in years who apparently hasn't aged a day).

Importantly, however, identification in this sense is not just the practice of determining sameness over *time*. Our perspectives on objects are constrained not only temporally, but also spatially. We always only perceive aspects of an object, and only at a certain scale. We know for a fact that physical objects have micro-structures, which we can't perceive in the same unproblematic sense as their macro-structures. Nevertheless, in an important sense, the macro-object just *is* the micro-structure, but from a different perspective. Think of Place's example of the strict identity relation between lightning and electrical discharges. The fact that the observable yellow flashes just are electrical discharges is something we needed to discover. So, my reports about what I'm seeing when I see lightning are reports of something that *happens to be* electrical discharges, in the same sense Smart once suggested that, "in so far as 'after-image' or 'ache' is a report of a process, it is a report of a process that *happens to be* a brain process." (Smart 1959: 144) And just as with determining sameness over time, also here, establishing the identity relation means acknowledging that, to the two (or more) distinguished relata do *not* correspond two or more objects with a unique spatiotemporal history. The fact that there are two or more relata here is not the result of there actually existing two or more different objects, but an inevitable consequence of the fact that our acquaintance with objects is always perspectival, in both space and time.[83] I will therefore refer to this first mode of identification as P-Identification, where 'P' stands for 'perspectival'. Here, the relevant concept of identity – which I will call P-Identity – refers to the strict identity relation of 'being identical to' or 'being one and the same thing'; in addition, I will use the phrase 'P-identical' to specify the relation between two objects that are strictly identical, e.g. "Hesperus is P-identical to Phosphorus". Indeed, it is the kind of identity at work within the classic identity theories, and the kind of identity that we've up till now been referring to as strict identity.

[83] The idea of taking sameness over time together with spatial sameness may strike some as unintuitive, yet it is explicitly endorsed by Smart, where he talks about objects as four-dimensional. See Smart 1959: 145.

In addition, note that P-identification does not require reference to abstract categories or types. We can identify an object as one and the same thing without knowing *what* it is, from a classificatory point of view. This is important, because it suggests that the identity theory need not be construed in terms of types and tokens. As I will further argue below, types and tokens are notions that are inseparable from our attributions of 'whatness' or C-identity.

4.3.2 C-Identification and C-Identity

Next to identifying something in the sense of determining 'one-and-the-sameness', we identify in the sense of determining *what* something is; that is, we identify something *as*, as opposed to identifying something *with*. And just as with P-identification, this happens mostly unreflectively; usually, we don't first see an object, only to infer afterwards what it is, for instance, a chair or a table. Rather, we ordinarily just 'see' these things immediately *as* chairs or tables. Sometimes, however, this dynamic of *seeing-as* gets interrupted, namely, when we are confronted with something of which we can't immediately tell what it is. It is on these intersections that this mode of identification reveals itself as something in which we have an active role to play. We never just see *objects*, we classify them, that is, we structure them against the background of a culturally shared classificatory scheme. I will therefore refer to this mode of identification as classification/categorization, or C-identification for short (where the 'C' can also be taken as indicative of the inherently cultural dimension of this form of identification). Within this mode of identification, the relevant concept of identity refers to *what* something is. I will call this classificatory/categorical identity, or C-identity for short. Also, I will use 'C-identical' to characterize the relation between objects with the same C-identity, e.g., Venus is C-identical with Mars, since both are C-identifiable as planets. For my present purposes, it is unnecessary to give a full analysis of this type of identification/identity, but it is worth highlighting a few characteristics that are directly relevant for our discussion concerning MR and identity theory.

C-identification, by its very nature, involves the determination of sameness in the sense of similarity. Class-membership is always a matter of being in some sense similar to the other members of the class. But similarity alone isn't sufficient. The similarity must have acquired the status of a classificatory criterion: it must have acquired a specific relevance in relation to our classificatory practices. This latter aspect deserves full emphasis, because it is especially here that it becomes clear that *what* something is, is never simply given. It always involves a normative element which entirely escapes us as long as we only consider the object and its properties. To put it in terms of discovery, although similarities are 'things' that can be discovered, the fact that some similarities have the status of a classificatory criterion, whereas others do not, is in itself not something that can be discovered in the same way[84] we discover similarities in the world. Consequently, we never simply discover the whatness of an object. This is what Wiggins is referring to when he speaks of the realist myth of the self-differentiating object. 'Whatness' or C-identity is something that only exists against a background of shared classificatory practices, in which certain similarities count as a relevant criterion, and others don't. This is not to introduce a kind of anything-goes relativism that would reduce classification to an arbitrary affair, but it does mean introducing a kind of relationalism that relates the 'whatness' of an object, and its identifiability as 'a cat' or 'a corkscrew' or 'an instance of multiple realization', to certain kinds of human practices. Notions like types, kinds, categories and classes are abstractions that cannot be made sense of in separation of our classificatory practices, and they are never simply 'given'. Abstract types (or kinds, categories, classes) are not themselves scientifically observable entities in the world. So any self-declared empirical theory that claims that particular entities are, as a matter of empirical fact, realizations of types, must eventually deal with the question of how these types – as well as their realizability – can be understood in empirical terms as well. If it cannot, the theory remains metaphysical and is no more empirical than a theory claiming that all particular beings are manifestations of one supreme being, or

[84] Of course, it can be discovered by, for instance, the cultural anthropologist who studies classification systems.

that all particulars are imperfect imitations of their perfect universal counterparts that exist in a separate intelligible realm. As we'll see below, arguing for the empirical nature of the MR thesis is precisely the task authors like Thomas Polger and Lawrence Shapiro have set themselves.

4.4 Applying the distinction

With the distinction between P- and C-Identity in the back of our minds, I now want to reconsider the idea that, as Polger rightly points out, "MR is at root a thesis about similarity and difference." (Polger 2009: 458) I first of all want to highlight the fact that both kinds of sameness or identity (P- and C-identity) are involved in the MR argument against identity. On the one hand, the idea of MR is inextricably linked to our classificatory practices or our acts of C-identification. It is, after all, a thesis about types. It says that one and the same type can be realized by relevantly different physical structures. However – and this is a pivotal move in my argument – I want to reconsider this definition of MR and suggest another, less metaphysically laden formulation of the MR thesis. Holding on to the realization relation, I propose that we redefine MR as the thesis that claims that some things that we classify as relevantly[85] the same are sometimes realized by physical structures that we classify as relevantly different. This is most clearly the case for functional objects like mousetraps, carburetors, corkscrews and computers, and it is supposedly also true for mental entities. Accepting the above reformulation means accepting that we connect, or rather *re*connect, the MR thesis to our classificatory practices. In light of what was said about the socio-cultural dimension of C-identification, I think this reformulation is warranted: Since C-identity (an object's 'whatness') is a notion that loses all meaning when we detach it from shared classificatory practices, and since it is precisely C-identity of which (multiple) realizability can be predicated, the MR thesis itself loses its meaning when we detach it from our classificatory practices.

[85] Of course, much hinges on what is to count as a relevant difference here. I will return to this issue extensively in section.

So inserting this element of classification into the definition of MR is, I take it, a legitimate move. [86]

Contrary to MR theory, identity theory involves first of all P-identity. When put in terms of our distinction between C- and P-identity, the general idea of the identity theorist is that, whenever we C-identify something as mental (a pain sensation, an after-image), we can in principle always P-identify it with something that we C-identify as physical (perhaps a neural state, or a brain-body-environment interaction). Now, the question whether the identity theorist's proposal is viable or even intelligible is one thing, but in light of the distinction between C-identity and P-identity, the idea that P-identity – the classic identity theorist's 'strict identity' – is ruled out because of the alleged multiple realizability of the mental is not at all obvious, for at least two reasons.

First, as we've just seen, MR is a thesis cast in terms of C-identity and C-identification: it says that certain things (in the most general sense of the term) that are C-identical (e.g., two tables) can be realized by physical structures that we C-identify as relevantly different (e.g., wooden vs. metal tables[87]). The thesis of MR, then, is inextricably tied to practices of *identifying as*. Identity theory, on the other hand, first of all involves identifying *with*. I say 'first of all', because to the extent that the identity theorist makes further claims about the kind of relata that are involved in the P-identity relation, (s)he also reverts to C-identification. Saying that pain is P-identical to neuronal firing is not only a claim about the kind of relation (P-identity), but also about the relata (C-identity). However, it would be a very uncharitable reading of Place, Feigl, Smart and other identity theorists if we would interpret their proposals as standing or falling by the accurateness of their specifications of the relata. In fact, none of the classic identity theorists seemed to be too concerned with the relata question. To my

[86] Until someone can come up with a naturalistically credible story that argues that objects do in fact wear their identities on their sleeves, regardless of our acts of classification, I suggest we accept the idea that the MR thesis can be defined as the thesis that things that we classify as relevantly the same are sometimes realized by physical structures that we classify as relevantly different.

[87] The example of the multiple realizability of tables is not accepted by everyone as a genuine example of MR. As we'll see in the next section, Lawrence Shapiro proposes a more restrictive criterion to distinguish actual from non-actual cases of MR.

knowledge, none of these authors ever suggested specific candidates for the identity relation. The idea that a classic identity theorist would ever have insisted on there being a strict identity between pain and C-fiber stimulation is a myth, which finds its origin in inaccurate representations of commentators. Neither Place, nor Feigl or Smart have ever insisted that pain is identical to C-fiber stimulation, yet some commentators keep suggesting that this is somehow a central tenet of classic identity theory. Roland Puccetti quotes William Lycan: "Consider the much touted version of the identity theory according to which pains or pain events are strictly identical with C-fiber stimulations" (Lycan 1974: 667). Puccetti rightly points out that Lycan's assertion

> is strange, because the attitude of most philosophers debating the mind-brain identity hypothesis has been that (a) nothing much hangs on what particular neural mechanism is supposed to be identical with a kind of psychological state or event, and (b) in any case the choice of a specific neural process is best left in the hands of neurophysiologists. (Puccetti 1977: 303)

Indeed, for the classic identity theorists, the key issue has always been to defend the relation, not the specific relata the identity relation is supposed to hold between. In fact, all classic identity theorists show clear restraint from any C-identification of the relata[88]. Nevertheless, most charges against the identity theorists' proposal, including Putnam's MR argument, seem to be aimed at precisely this uncharitable interpretation[89].

The second point I want to highlight is that the MR thesis is – as we've already seen – not intrinsically a thesis about the physical-mental relation, as it clearly also applies to ordinary objects like corkscrews (Shapiro's toy example, as we'll

[88] Take Feigl, for instance: "Just which phenomenal qualities correspond to which cortical-process patterns has to be determined by empirical investigation." (Feigl 1958: 42)

[89] In their recent *The Multiple Realization Book*, Polger and Shapiro make a similar complaint in response to Putnam, who claims that "the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can be clearly applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain state theory has collapsed. It seems to me overwhelmingly probable that we can do this. (1967 in 1975: 436–7) Polger and Shapiro are right in thinking that "this presentation of the burden of proof on identity theory rests on an uncharitable interpretation of both historical and contemporary identity theorists." (Polger & Shapiro 2017: 34)

see). As I've already indicated above, all authors invoke non-mental entities as examples of MR to clarify the thesis in regard to mental entities. So if we accept the above reformulation of MR as a thesis about mental entities, we should also accept the following reformulation of MR as a thesis about things in general, regardless of their being mental or not: MR is a thesis that says that some things that we classify as relevantly the same, can be realized by physical structures that we classify as relevantly different. Whether these 'things' are supposed to be mental entities like pains or after-images, or functionally defined non-mental things like mousetraps or corkscrews makes no difference for the thesis' applicability.

4.5 MR as orthogonal to identity theory

When we take a closer look at these two points, it should become clear that the idea of MR entailing an argument against identity theory is problematic from the start. With regard to the first point, we should ask how a thesis that is essentially about C-identity relations can make any claims about possible P-identifications. It is perfectly possible to have different C-identifications of potentially different things where these potentially different things can nevertheless be P-identified as one and the same thing. Take the classic 'Water = H_2O' example. What this says is that something we C-identify as water, and at the same time C-identify differently as a collection of hydrogen-oxygen bonds, is one and the same thing (P-identical). Now, whatever other identity issues this example may have, there does not appear to be a special conflict between both kinds of identification. In general, it is unclear how in itself, C-identities could ever be taken to endorse or rule out possible P-identities. Issues of C-identity just seem to be orthogonal to issues of P-identity. Of course, if it turns out, for example, that creatures without a brain can also experience pain, then the identity theorist would be wrong in P-identifying pain with neuronal firing, as the chemist would be wrong in P-identifying water with C_2H_6O. But this in no way means the proponent of psychophysical identity must be wrong in thinking that the relation holding

between a mental and a physical entity is P-identity. It just means that the identity theorist can be mistaken about which mental entities are identifiable with which physical entities. But this has always been acknowledged by identity theorists. To put it bluntly, the identity theory does not logically exclude the possibility of there being two particular occurrences of pain that are each P-identical with two very different physical states or processes. In addition, it should be noted that, despite the MR theorist's question-begging and unsupported assumption that pain can be realized in completely different physical substrates, the empirical evidence point in the opposite direction. Without wanting to make a case for neuro-centrism, the fact that Putnam's original example includes creatures that all have brains seems to point to a relevant *similarity*, rather than a difference. In any case, the burden of proof still lies with the MR loyalist to show that different creatures that can be attributed with the same mental properties are really *relevantly* different in a physical respect.

With regard to the second point, it becomes even more apparent how the MR thesis is unequipped to counter the idea of mental-physical identity. If the mere fact of some entity's multiple realizability would entail an argument against P-identity, then *this should be reflected in all instances of multiple realization, not just in the multiple realizability of the mental*. So, for instance, suppose we agree with Shapiro (2000) and accept that something counts as multiply realizable if it can bring about the same function in causally different ways. His example, as already mentioned, is that of the type 'corkscrew'. 'Corkscrewness' is multiply realizable because different corkscrews can bring about the uncorking of a bottle in causally different ways. Waiter's corkscrews uncork bottles in an apparently relevantly different way than winged corkscrews. In other words, suppose we accept Shapiro's stipulation – and it is nothing *but* a stipulation – that two or more C-identical objects (two or more particular corkscrews) can nonetheless be C-identified as relevantly different (a waiter's corkscrew vs. a winged corkscrew) because of the accepted criterion of 'bringing about the function in causally distinct ways'. Are we now to conclude that, because of the stipulated multiple

realizability of 'corkscrew', corkscrews can no longer be P-identical with physical structures? Of course, we could again be wrong in C-identifying the physical structure, for instance, if we would claim that corkscrews are strictly identical with structures of H_2O molecules, but this doesn't mean that we would have to give up on the idea that corkscrews *just are* physical structures.

The same point can be made with regard to the alleged multiple realizability of the functional in general. When we take a closer look at the ways human beings classify the world, we see that, for the C-identity of some objects, matter is (relatively) irrelevant, whereas for others, it *is* relevant for their classification. To put it very simply: actual trees can't be made from metal, but for something to be C-identifiable as a table, it doesn't matter whether it is made from wood or metal. This is because tables, like other functional objects, are classified according to a functional criterion (say, allowing to be seated and have dinner at), which itself allows for a *relatively* wide material variety (tables can't, of course, be made from helium gas). Now, we *could* say that this means that the type 'table', like other functional types, is multiple realizable in the physical. The point, however, is that we don't *have* to say this. We may just as well say that we simply classify things according to different relevancies and avoid altogether talk of types like 'tableness' or 'corkscrewness', as well as invoking notions like 'realization', 'implementation', or any other kind of metaphysical relation that is supposed to relate abstract types to concrete particulars. In one context, one and the same object can be classified as a table, in which case it doesn't matter whether it is made from wood or metal; but in another context, the exact same object can be classified as firewood, in which case it matters a lot that the 'table' is made from wood. But outside these classificatory contexts, the object (the table/firewood) *just is* its physical spatiotemporal structure, which is all the identity theorist wants to say about mental 'objects' too. Perhaps it will be objected that the analogy with functional objects doesn't hold for the mental-physical case. This depends on one's preferred metaphysical theory of the mental. If one is a convinced Cartesian dualist, then of course the above comparison fails. Crucially,

however, it seems that precisely the functionalist, who accepts the mental as a functional category, can't object to the comparison.

4.6 MR as an empirical thesis

To properly evaluate the argument from MR against identity, it is also necessary to examine in more detail what it means for MR to be an empirical thesis. After all, identity theory has always been put forward as an empirical theory, so for MR to be a better alternative, it must obviously have the status of an empirical thesis as well. But what does it mean, exactly, to say that MR is an empirical thesis? Polger warns us that a certain reading of the MR argument might turn it into nothing more than a metaphysical stipulation, incompatible with strict identity on a priori grounds alone. Saying that a mental kind Ψ cannot be identical to a physical kind Φ because of Ψ's multiple realizability begs the question entirely. In addition, recall that on this reading, the assumed multiplicity of the realizations becomes entirely redundant. P-identity is already ruled out by the stipulated realization relation alone. It doesn't matter whether there are *multiple* realizations or not. However, since authors like Polger and Shapiro want to hold on to Putnam's suggestion that MR is to be understood as an empirical thesis for which evidence can be cited, these authors conclude that we should assume that the question-begging reading simply can't be right. Polger and Shapiro (2016) go to great lengths to come up with an interpretation of MR that secures its status as a plausible empirical theory. Below, I will turn to these authors' evaluation of MR as an empirical thesis. First, however, a short recapitulation.

MR is a thesis inseparable from acts of classification or C-identification. It claims that particular mental entities that are classified as belonging to a single type Ψ can be realized by particular physical entities that are classified as belonging to different types Φ_1, Φ_2...Φ_n. MR presupposes acts of C-identification in that the relata of the realization relation must already be C-identified. With regard to the realizable type: you can't say of something that it is multiply realizable if that something has not already been identified as this or that kind of thing. It is of a

thing's *whatness* (its C-identity) that multiple realizability is predicated. Yet, as said, determining C-identities is always also a socio-normative affair. Similarity based classification always also involves classificatory criteria, which are socio-normative entities. Nothing has in and of itself the status of a classificatory criterion, which is to say, there are no kinds outside our classificatory practices, natural or not. Yet, despite Wiggins' admonition, the idea that determining C-identity is an entirely objective affair, i.e., a matter of discovering some hidden essence which by its very nature "announces itself as the very object it is", is still wide-spread. Within discussions of MR and identity theory, the idea manifests itself as *type-realism*, meaning here that types are conceived of as entities with an existence outside our classificatory practices. One particularly relevant example of type-realism comes from Lawrence Shapiro's work on MR.

4.6.1 Shapiro's criterion

Just as we can ask questions about the C-identification of corkscrews, carburetors or planets (should Pluto be C-identified as a planet?), Shapiro asks himself a similar question in relation to multiple realization itself:

> Before it is possible to evaluate the force of MRT (*Multiple Realization Thesis*)..., we must be in a position to say with assurance what the satisfaction conditions for MRT *actually* are. (Shapiro 2000: 636; first italics added; second italics mine)

But contrary to what Shapiro seems to believe, these satisfaction conditions are not things that are – and here we see Wiggins' realist myth at work – "actually" there, and for which evidence can be cited. In his already mentioned example, Shapiro holds that two waiter's corkscrews do not count as two different realizations of the type 'corkscrew'. Yet, 'corkscrewness' *does* still count as a multiply realizable kind because, next to waiter's corkscrews, we also have winged corkscrews, which also remove corks from bottles, but in a different "causally relevant way" (Shapiro 2000: 645): "To say that a kind is multiply realizable is to say that there are different ways to bring about the function that defines the kind." (Shapiro 2000: 644) And with regard to the corkscrew example,

he adds: "The moral of this example is that multiple realizations count *truly* as *multiple* realizations when they differ in causally relevant properties …" (Shapiro 2000: 644; first emphasis mine) Although Shapiro presents this criterion as a means to provide more insight in MR as "a thesis about the physical world" (Shapiro 2000: 637), in light of what has been said above about the non-discoverability of classificatory criteria, Shapiro's so-called "recipe" (Polger & Shapiro 2016) for distinguishing "genuine cases of multiple realizability" (Shapiro 2000: 636) from non-genuine ones, and for determining whether some kind "does *truly* admit of multiple realization" (Shapiro 2000: 652; my emphasis), is nothing but a stipulation. For there is nothing we can *truly* say about the physical world that would make Shapiro's criterion of causal distinctness *the true* criterion of picking out "actual", "genuine" or "true" cases of multiple realization from non-actual, non-genuine or untrue cases (these latter cases which, it should be added, for Shapiro include Putnam's original examples, which were supposed to make the whole idea of MR intuitive to begin with). Thinking that Shapiro's criterion somehow contributes to a better understanding of how the physical world really is, is the same as saying that, because we have a criterion for classifying planets which entails that Pluto is no longer a planet, we now have discovered something new about our solar system. Adopting the causal distinctness criterion as a way of distinguishing genuine from non-genuine cases of MR does not, and *cannot* offer us any new insights about the way the physical world really is; the causal distinctness criterion's only relevance is that, without such a criterion, MR might collapse into something too ubiquitous and potentially uninteresting. If, for instance, difference in material composition were to count as a criterion for MR (as it seemed to do for Putnam), so that some type can be said to be multiply realizable if its tokens can be of distinct material types (say, a wooden vs a metal table), then way too many objects would have to count as instances of multiply realizable types, pretty much in the same way that, were we to still count Pluto as a planet, our solar system would soon be counting a lot more than 9 planets (there are now believed to be many other Kuiper Belt Objects that would have at least as much the right to be classified as a planet than Pluto). All Shapiro's criterion does is stipulate which entities can be labelled as instances of MR, and

which cannot. And indeed, this is all we can expect from a stipulated classificatory criterion. However, with regard to the MR-identity discussion, it should be noted that we have at least one good reason for *not* wanting to adopt Shapiro's criterion. The reason is that the causal distinctness criterion already assumes that MR only applies to causal phenomena (like the uncorking of a wine bottle). So, to assume that the MR thesis even applies to mental entities requires assuming that the relation between the physical and the mental is causal. In other words, accepting Shapiro's criterion for MR only makes sense against the background assumption that mental entities are somehow being brought about by physical entities. But whether the same mental entities are brought about in causally different ways makes no difference for the identity theory. Merely assuming that the mental is brought about by the physical (regardless of whether MR is true or not) is enough to rule out strict identity by default. Once again – as was the case with the deductive a priori argument from MR against identity – we must already accept *a priori* that the identity theorist is wrong. After all, strict identity relations do not reduce to causal relations. There is no causality involved in being one and the same thing. So, as long as we accept that the mental is caused or "brought about" by the physical, the identity theorist doesn't stand a chance. It is remarkable, then, that we find this assumption with authors (Polger & Shapiro) who claim to endorse an identity theory.

In any case, the question still is: even if we accept Shapiro's "recipe" for distinguishing genuine from non-genuine cases of MR, how does this help to make the MR thesis acceptable as an *empirical* thesis? On Shapiro's reading, the empirical status of the thesis is already presupposed. Again, all Shapiro's criterion does is saying that, if we can empirically establish that two or more numerically different entities bring about the same effect in a relevantly different causal way, then we can genuinely classify these entities as belonging to the 'multiply realizable type' type. But this already bypasses the more fundamental issue of how an essentially metaphysical thesis can be rendered in an empirically falsifiable hypothesis for which evidence can be cited. Shapiro's interpretation of the MR thesis remains silently committed to a metaphysical picture in which

particulars are understood as token realizations of types. His ambition is to determine when something can be said to be *multiply* realized, which already assumes the metaphysical picture in which particular objects (e.g., corkscrews) are to be understood as realizations of a type (e.g., 'corkscrewness').[90] But this metaphysical picture is not itself empirically motivated. Applied to Shapiro's criterion for MR, observing that certain effects can be brought about in causally different ways in no way requires us to accept a metaphysics in which the particulars involved are to be understood as realizations of types. On the other hand, however, it seems that there can only be a real conflict between identity theory and MR if it is assumed that the identity theorist is committed to this exact same metaphysical picture. After all, if the fact that one and the same mental type is realized by different physical types is supposed to be in direct conflict with identity theory, this means that we should understand identity theory as making the opposite claim, namely, that one and the same mental type is realized by one and the same physical type. In other words, for *multiple* realization to be a threat for identity theory, we must understand the latter as endorsing *single* realization. It is unclear, however, why the identity theorist should commit herself to this metaphysical picture in which particular entities are understood as token realizations of some abstract type. Yet, it is only by framing the identity theory in this way that MR can present itself as being incompatible with identity theory.

4.6.2 Polger's strengthening manoeuvre

According to Polger, to understand MR as more than a metaphysical stipulation, we need to make a distinction between a metaphysical reading on the one hand, and an ordinary reading on the other. It is on this second reading that MR can lay

[90] Although Shapiro and Polger are critical about how to understand the realization relation, they do hold on to the general idea that at least some objects (functional objects) are realizations of types or kinds. See, for instance, Polger & Shapiro 2016, especially chapter 2. Here, they write: "[I]t seems more convenient to us to speak about particular objects being, or not being, corkscrews; that is, falling under the kind *corkscrew*. This looks like a case of objects realizing kinds." (Polger & Shapiro 2016: 27)

claim to being an empirical hypothesis. Polger puts the ordinary interpretation of 'realization' in terms of 'mediation', which he defines as follows:

> Mediation is a generic relation, maybe stronger than mere correlation or association and involving some spatio-temporal correspondence. Mediation is the kind of correlative relation that might plausibly be empirically observed. (Polger 2013: 869)

If we ignore, for the sake of the argument, that this vague definition runs itself the risk of being accused of begging the question, on this reading in terms of mediation, the charge that the argument from MR against identity is question-begging with regard to the realization relation is no longer warranted. Interchanging realization for the allegedly empirically observable mediation relation, Polger schematizes MR as follows:

> Ψ is mediated by physical state P_1 [in creature C_1.]
>
> Ψ is mediated by physical state P_2 [in creature C_2.] (Polger 2013: 869)

These premises presumably entail an argument against identity. According to Polger, instead of a deductively valid, but uninteresting argument, we now have a potentially interesting empirical likelihood argument. What remains of the MR thesis is the claim that it is empirically verifiable that one and the same psychological type is mediated by multiple, relevantly different physical properties, states or processes in different creatures. This is supposed to make psychophysical identity unlikely.

Polger holds that MR "is most plausibly thought of as the claim that psychological state kinds are shared in common across at least some physical creature kinds, for example, across species." (Polger 2013: 870) But what, precisely, does this claim implicate? If all there is to MR theory is the claim that different creatures can have the same kind of mental states, e.g., that both cats, dogs and humans can be said to feel pain, then the MR claim becomes fairly trivial, and we should certainly wonder why the truth of this claim, which everyone probably will accept, is thought to pose such a threat for the identity

theorist. It is unlikely that identity theorists like Place, Feigl and Smart wanted to deny that different animal species are capable of experiencing pain. So if we want to keep taking the debate seriously, this is probably not how we should understand the MR claim. What makes this reading of the MR claim trivial is that it leaves out some essential elements of the MR thesis. What makes the MR claim more than an expression of the common sense idea that different creatures can experience the same kind of mental state, e.g., pain, lies in the way MR attributes sameness to types, as well as the specific realization relation it postulates as holding between a single type and its multiple realizations. The alleged sameness of the types is the cornerstone of the deductive MR argument against identity. Yet, it is precisely at its cornerstone that the thesis is at its most vulnerable, or so I'll argue below.

4.7 Sameness in relation to types

In their recent *The Multiple Realization Book*, Polger and Shapiro claim that "multiple realization requires that the psychological functions be of *exactly the same kind*" (Polger & Shapiro 2016: 50, m.e.). In the following, I want to examine more closely what it means when authors attribute 'exact sameness' to types. To avoid misunderstandings, I want to make it as clear as possible that I will *not* be talking about the issue of determining whether or not two mental states can be classified as being of the same mental type (or, in our terminology, whether these states are C-identical), for instance, whether what an octopus is feeling when it is being stabbed is classifiable as the same kind of sensation a stabbed human being experiences, i.e., pain. This is an issue for everyone, not just the MR theorist. I also won't be addressing the difficult, though important question of whether the idea of strict identity even applies to mental entities (is it sensible to ask whether my toothache-sensation now is P-identical with my toothache-sensation from five minutes ago?) What I *do* want to focus on is the question of how we are to understand 'exact sameness' in relation to types because, as we've seen above, the idea of sameness allows for two very different interpretations, depending on the

mode of identification (C- or P-identification). For a proper evaluation of the MR argument against identity, it is important that we are clear on what kind of 'exact sameness' is at work here. That this 'exact sameness of type' is essential for MR to form a potential threat for the identity theorist is explicitly acknowledged by Polger:

> For multiple realizability to be a problem for identity theory it is not sufficient that some wildly different creature have some conscious state or other; it must be that different creatures can have *exactly the same*—empathetic—*kinds* of mental states. (Polger 2013: 13; my emphasis)

Yet despite Polger's explicit acknowledgement of the identity requirement for types, the inherent ambiguity within the phrase 'exactly the same kind' stays unnoticed. To be sure, this phrase is commonplace in the literature on MR. Consider the following standard formulations of MR:

> As far as anyone knows, different organisms are often in psychological states of *exactly the same type* at one time or another, and a given organism is often in psychological states of *exactly the same type* at different times. (Block & Fodor 1972: 159)

> The multiple realizability thesis contends that *a single mental kind* (property, state, event) can be realized by many distinct physical kinds. (Bickle: 2016, m.e.) [91]

With regard to the question of what kind of sameness is at play here, it seems rather obvious that in the above formulations, what is predicated of types is 'one-and-the-sameness', or P-identity. The type of which multiple realizability is predicated must be strictly, or numerically, or P-identical across the different realizations. Furthermore, it is the *type* of which strict identity is predicated, not the particular instantiations (actual states or processes). Pointing this out may seem unnecessary, but it is something which needs to be kept in mind to avoid confusion. Consider for instance the following alternative formulation of the MR thesis:

[91] Bickle, John, "Multiple Realizability", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/multiple-realizability/>

> The claim of multiple realizability is the claim that *the same psychological state* can be realized by different brain states. (Bechtel & Mundale 1999: 176, m.e.)

In this formulation, the authors appear to be attributing multiple realizability to a given psychological state, rather than a psychological state type. If we assume that they are, in fact, talking about one and the same instance of some psychological state type, instead of the type itself, MR is taken to be saying that two or more different creatures with brains (say, me and my cat) can both experience one and the same, P-identical psychological state (say, one and the same toothache), perhaps at the same time. I don't think anyone takes this idea seriously. To be clear, saying that two creatures can experience exactly the same sensation in the sense of P-identity does not mean they each have a sensation which is very similar to one another, for this would be confusing our P-identity attributions for C-identity judgments. 'Exact sameness of sensation' in the sense of P-identity means that there is only one sensation which both creatures experience, which is fairly implausible. However, in the literature, the fundamental distinction between both kinds of sameness tends to get blurred, and in Polger's writings as well. On the one hand, as we've seen, he acknowledges the importance of the strict identity of the realizable types if MR wants to be an argument against identity theory. At the same time, however, we see that he himself confuses strict identity with mere similarity. Consider this passage:

> Despite what Bill Clinton may say, none of us believes that he truly felt our pain. We expect that there are differences between individuals, even within individuals over time, that make it unlikely that one sensation is ever exactly similar to another. (Polger 2013: 13)

Here, by invoking the idea of exact similarity, Polger conflates P-identity with C-identity. I agree with the first sentence, which claims that no one believes that two or more creatures can ever experience exactly one strictly identical sensation. But strict identity does not equate to a relation of similarity, which is what is being implied in the second sentence. The essential difference is that, with P-identity, there is in fact only one entity involved, whereas in Polger's example, the exact similarity relation is conceived of as potentially holding between two

entities, i.e., two sensations. Simply put: P-identity does not equate to similarity, not even exact similarity (if such a thing even exists).

More importantly, on closer inspection, this conflation of P-identity and C-identity is also at work in Polger's version of the argument from MR against identity. Recall:

1. Ψ is mediated by physical property, state or process P_1 in creature C_1

2. Ψ is mediated by physical property, state or process P_2 in creature C_2

C: Ψ is not identical with P_1 or P_2

For the argument to be deductively valid, the Ψ's occurring in the premises and conclusion must, as a formal requirement, of course refer to one and the same entity. This is what Polger means when he says that, for the MR argument to work against identity, the multiple realizability must be predicated of *exactly* the same type. However, resolving the issue of begging the question with regard to the realization relation is not enough to turn the MR thesis into a fully-fledged empirical claim which can be used in an argument against identity. The problem is that Polger's strengthening manoeuvre makes it impossible to continue attributing strict identity to mental types, because the mediation relation, unlike the realizability relation*, is a relation that is defined as holding, not between types and their physical realizations, but between actual instances of mental and physical states*. Reconsider Polger's definition of mediation:

> Mediation is a generic relation, maybe stronger than mere correlation or association and involving *some spatio-temporal correspondence*. Mediation is the kind of *correlative relation* that might plausibly be empirically observed. (Polger 2013: 869; my emphases)

On this construal, mediation can never be a relation holding between realizable types and physical realizers because types simply aren't spatiotemporal entities that can correlate with things. Rather than referring to a single type, Polger can only be referring to particular instances of a single type. But then this should be reflected in his formulation. So instead of writing:

Ψ is mediated by physical state P_1 [in creature C_1]

Ψ is mediated by physical state P_2 [in creature C_2]

Polger should have written:

ψ_1 is mediated by physical state P_1 [in creature C_1][92]

ψ_2 is mediated by physical state P_2 [in creature C_2]

This adjustment from types to tokens makes the above unsuitable as premises in a deductive argument against identity because here, the required P-identity of the type collapses into C-identity between different instantiations of the type. ψ_1 and ψ_2 do not pick out one and the same entity, but two numerically different entities between which only a relation of relative similarity can hold. But, as should be clear, relative similarity is not enough for a deductive argument against identity. Only strict identity will do. So again, the required 'exact sameness' relation (P-identity) collapses into a relative sameness connected to our classificatory practices (C-identity). Apparently, substituting the metaphysical realization relation for an allegedly empirical relation like mediation requires us to jump from the metaphysical level of types to that of actual tokens. But this also means that MR is no longer suitable to figure in a deductive argument against identity.

We could ask whether Polger's account is still an account of MR at all, since here, the realization relation can no longer be understood as a one-to-many relation, but only as a many-to-many relation between relevantly similar mental tokens and relevantly different physical tokens, so that we could say that

ψ_1 is realized by α, ψ_2 is realized by β... ψ_n is realized by ω,

where $\psi_{1,2...n}$ stand for different mental tokens of the same type Ψ and α, β and ω stand for different physical tokens of different types A, B,... Ω. What this says is that individual mental entities (numerically different toothaches, say) that we C-identify as belonging to the same mental category (the 'toothache' category) can,

[92] 'ψ' standing for an actual instance of a mental entity, not the mental entity type.

as a matter of empirical fact, be realized by physical structures that we classify as belonging to different physical categories. But if this really is what MR loyalists have in mind when they claim that psychological types are, *as a matter of empirical fact*, multiply realizable, then the identity theorist shouldn't worry too much, for reasons that will be presented in the following summarizing conclusion.

4.8 Summary and concluding remarks

1. On Polger's empirical reading, there is no deductive argument from MR against identity, because the required 'exact sameness of type' in the sense of P-identity stays unsatisfied. The expression 'being realizations of one and the same type' is still a metaphysical, not an empirical one. This being said, there is, however, a sense in which it is perfectly unproblematic to speak of 'one and the same type', a sense that does not commit us to a metaphysical picture that postulates types as the kind of things that can be attributed with strict identity and multiple realizability. When we ordinarily say that a waiter's corkscrew and a winged corkscrew belong to 'exactly the same type', all we are asserting is that both objects can be properly C-identified as corkscrews and that they can be *labelled* exactly the same. The crucial point, however, is that the 'exact sameness' is now predicated of our classificatory practices, not of some type or category believed to exist outside these practices. In other words, correctly saying that two objects belong in the same category does not require the existence of a third object (the category or type), over and above the two objects. As I've indicated earlier, the need to invoke metaphysical entities like types simply disappears when we bring our classificatory practices back into view. Saying that two corkscrews are realizations of 'exactly the same type' can be reformulated as saying that we can properly classify these two objects as corkscrews because they satisfy a socio-culturally shared classificatory criterion ('having the purpose of removing corks

in a certain way').[93] So the metaphysical assumption that there are types of which both P-identity and multiple realizability can be predicated is empirically unwarranted as well as unnecessary once we take our actual classificatory practices into account. But the standard deductive argument from MR against identity is entirely based on this metaphysical assumption.

2. On the empirical reading, the MR thesis runs the risk of collapsing into the trivial claim that physically different creatures can have similar psychological experiences. To avoid triviality, the MR theorist must solve at least the following two problems. First, it must be made clear of how we are to determine empirically when certain similarities and differences are to count as *relevantly* similar/different. Apparently, Putnam once believed that humans, reptiles and mollusks are such remarkably different physical beings as to rule out 'brain-pain' identities. However, as said, the fact that these different creatures are all C-identifiable as 'brained organisms' may just as well be taken to point to a relevant *similarity*, rather than a difference. Moreover, even if we would have an agreed upon criterion to determine what is to count as relevantly similar/different *today* (which we don't), and even if according to this criterion, some organisms are properly C-identifiable as relevantly similar in a psychological, yet relevantly different in a physical sense, this doesn't mean that we will never have to revise this criterion. This point has already been made long ago by Jaegwon Kim, Paul Churchland and others[94]. In any case, as things stand, we are still far removed from robots with toothaches. Second, and perhaps more importantly, to 'de-trivialize' MR, the MR theorist owes us an account of the nature of the realization relation that is now – on the empirical reading – said to hold between a realized psychological state and a realizing physical state. In this respect, the identity

[93] To be clear, denying that there are types in nature does not entail denying that nature is structured in certain non-random ways. It entails the denial that nature *classifies* itself. And it is only within a context of classificatory practices (C-identifications) that ideas like types begin to make sense.

[94] As early as 1972, Jaegwon Kim writes: [T]he mere fact that the physical bases of two nervous systems are different in material composition or physical organization with respect to a certain scheme of classification does not entail that they cannot be in the same physical state with respect to a different scheme. (Kim 1972, in Block 1980: 234–235) And a few years later, Paul Churchland writes: It is entirely possible , for example, that we, the gaseous Nebularians, the crystalline Plutonians, and any other persons lying about are all 'super-heterodyning negentropy flowers', where this expression is a part of the vocabulary of some future and more fully articulated version of statistical thermodynamics, a theory of awesome generality as it stands. (Churchland 1979: 112)

theorist holds the advantage. According to him, the relevant relation is strict identity, which is a perfectly unproblematic empirical relation. But how are we to understand 'realization' in an empirically respectable sense? Since so much hinges on realization, Polger notes that

> [y]ou might therefore expect that a great deal has been done to clarify the realization relation. But you would be wrong. …Once in a while it is noticed that realization is in need of scrutiny, but almost invariably that is left as a project for another day. (Polger 2007: 234)

And to this day, we are left with as many interpretations of the realization relation as there are authors relying on the notion. Quoting again from Polger:

> Wilson, for example, wants a relation that covers not only mental states but also cases of realization in "the banking system, the criminal justice system, or the electoral system" (2001: 14). Likewise, Poland's account potentially includes "social and cultural objects" (1994: 67). Heil (1992) mentions the realization of mental states, but also the realization of a desk by its parts. And Gillett's (2002) primary example of realization is not that of mental states but rather the hardness of a diamond. (Polger 2004: 118)

However, the problem isn't only that there is no agreement on how to understand the realization as a metaphysical relation, the more pressing issue is that, even given such a consensus, it is still unclear how to make sense of this metaphysical relation in an empirically interesting, i.e., explanatory, sense, such that it rules out identity, and at the same time avoids a relapse into some form of dualism. How does an actually occurring physical state manage to realize an actually occurring mental state? These are questions the identity theorist doesn't have to lose sleep over.

3. On the empirical account, the MR argument against identity collapses into the idea that two mental states that we classify as relevantly and relatively the same can be realized/mediated by two or more physical structures that we classify as relevantly different. This does not rule out identity deductively, but it is supposed to make it 'unlikely'. But why? To be absolutely clear: logically speaking, a

psychophysical identity theory does *not* require that, every time we identify two mental entities as relatively/relevantly similar (C-identical), we must, in principle, always also be able to discover two physical entities which we can also identify as relatively/relevantly similar. Again, *strict identity does not equate to similarity*. Otherwise put: being P-identical does not equate to being C-identical. What the identity theorist maintains is that, whenever we C-identify something *as mental*, we will always also be able to C-identify it as *physical* (perhaps neural), because they are strictly identical (P-identical). Crucially, however, this does not entail that two relatively/relevantly similar mental entities could not each be P-identifiable with two physical structures that we would C-identify as relevantly different. There is no inconsistency in the idea that tokens of a single mental type (pain, say), could be strictly identical with tokens of different physical types (organic vs inorganic structures, for instance). There is, in other words, nothing incoherent about the idea of a *multiple identity thesis*. Moreover, when we approach matters from our different identificatory practices (C- and P-identifications), we see that the possibility of 'multiple identities' is completely compatible with functionalism. The functionalist holds that two or more physically different structures can nonetheless be functionally identical. In terms of our classificatory practices, this comes down to saying that two structures that we C-identify as physically different according to some criterion A can simultaneously be C-identified as the same kind of thing according to some functional criterion B. Nevertheless, these functional entities are, from the perspective of P-identification, strictly identical with their physical structures. Functionally classified objects like chairs, mousetraps, carburetors or corkscrews can have relevantly different physical materializations, yet these objects and their physical structure are not two different things: they are one and the same thing, in the sense of P-identity. *The difference is a result of our different classifications, not of some real difference 'out there'*. One could argue that the analogy does not hold when it comes to psychophysical relations, but this minimally requires denying that psychological states are functional states. Ironically, it is precisely the functionalist who can't deny this. Moreover, to see just how compatible

classic identity theory with functionalism and MR is supposed to be, we need only remind us of the following passage in Smart:

> If the brain-process theory is correct, then it is in principle possible that an appropriately constructed robot might be conscious *i.e.* have sensations. If in its (perhaps electronic) brain there were the right sort of processes, analogous to those that go on in us when we are conscious, then this robot would be conscious too. (Smart 1963: 105)

References

Bechtel, W. & Mundale, J. (1999). Multiple realizability revisited: linking cognitive and neural states. *Philosophy of Science 66*: 175–207.

Bickle, John, (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Block, N. & Fodor, J. (1972). What psychological states are not. *Philosophical Review 81*: 159-81.

Block, N. (ed.) (1980) *Readings in Philosophy of Psychology 1*. Cambridge, Mass.: Harvard University Press.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. New York: Cambridge University Press.

Evans-Pritchard, E.E. *Nuer Religion*. Clarendon Press: Oxford. 1956

Feigl, H. (1958). The "Mental" and the "Physical"', in H. Feigl, M. Scriven and G. Maxwell (eds.), *Concepts, Theories and the Mind-Body Problem* (Minnesota Studies in the Philosophy of Science, Volume 2), Minneapolis: University of Minnesota Press; reprinted with a Postscript in Feigl 1967.

Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis 64*: 316–323.

Heil, J. (1992). *The Nature of True Minds*. New York: Cambridge University Press.

Hutto, D.D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA.: MIT Press.

Kim, J. (1972). Phenomenal properties, psychophysical laws, and identity theory. *Monist 56*: 177–192. Excerpted in Block (1980) under the title "Physicalism and the multiple realizability of mental states."

Kirchhoff, M. & Hutto, D.D. (2016). Never mind the gap: neurophenomenology, radical enactivism and the Hard Problem of Consciousness. *Constructivist Foundations 11*: 302-309.

Lycan, W. (1974). Kripke and the materialists. *Journal of Philosophy 71*: 667-689.

Myin, E. & Loughlin, V.(in press). Sensorimotor and enactive approaches to consciousness, in Gennaro, R. (ed.) *Routledge Handbook of Consciousness*.

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

Papineau, D. (1998). Mind the gap. *Philosophical Perspectives*, *12*: 373-89.

Place, U.T. (1956). Is consciousness a brain process? *British Journal of Psychology 47*: 44–50.

Poland, J. (1994). *Physicalism: The Philosophical Foundations*. New York: Oxford University Press.

Polger, T. (2002). Putnam's intuition. *Philosophical Studies 109*: 143–170.

Polger, T. (2004). *Natural Minds*. Cambridge, MA: MIT Press.

Polger, T. (2007). Realization and the metaphysics of mind. *Australasian Journal of Philosophy 85*: 233–259.

Polger, T. (2009). Evaluating the evidence for multiple realization. *Synthese 167*: 457–472.

Polger, T., 2013, 'Realization and Multiple Realization, Chicken and Egg.' *European Journal of Philosophy*, 23 (4): 862-877.

Polger, T. & Shapiro, L. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press.

Puccetti, R. (1977). The great C-fiber myth: a critical note. *Philosophy of Science 44*: 303-305.

Putnam, H. (1975). *Mind, Language, and Reality: Philosophical Papers*, Volume 2. Cambridge:
Cambridge University Press.

Shapiro, L. (2000). Multiple realizations. *Journal of Philosophy 97*: 635–654.

Shapiro, L. (2004). *The Mind Incarnate*. Cambridge, MA: MIT Press.

Shapiro, L. (2008). How to test for multiple realization. *Philosophy of Science 75*: 514–525.

Smart, J.J.C. (1959). Sensations and brain processes. *Philosophical Review 68*: 141–156.

Smart, J.J.C. (1963). *Philosophy and Scientific Realism*, London: Routledge & Kegan Paul Ltd.

Wiggins, D. (2001) *Sameness and Substance Renewed*, Cambridge: Cambridge University

Press.

Wilson, R. 2001. Two views of realization. *Philosophical Studies 104*: 1–30.

5 Identity Reconsidered: taking a dual perspective on the Hard Problem of Consciousness

Abstract

Despite functionalism's long reign in philosophy of mind, it has never fully managed to carry off the older idea that the mind-matter relation might be a relation, not of multiple realizability, but of strict identity. Nowadays, we see a resurgence of identity-theoretical proposals in the so-called E-approaches to cognition, and especially in enactive and radical enactive approaches. Here, it is claimed that assuming a strict identity between certain physical structures and phenomenal consciousness isn't merely a viable option, it is perhaps the only way to avoid the Hard Problem of Consciousness. This paper wants to argue that the Hard Problem of Consciousness is a pseudo-problem that should indeed be avoided, rather than solved, and that this can be done by adopting a specific version of identity theory, one which isn't neuro-centric and which also avoids collapsing into ontological reductionism. This version of identity theory is based on classic work by Herbert Feigl, who provides one of the most elaborated, yet at the same time most overlooked identity theories. Inspired by his work, I will defend, what I will call, a dual perspective theory. The theory will be contrasted with, on the one hand, neuro-centric and reductionist identity theories, and, on the other hand, with other mind-body relation proposals such as supervenience, neutral monism and dual aspect theory. To explain the idea of 'dual perspectives', I shall rely on some of Merleau-Ponty's phenomenological insights.

5

Identity Reconsidered: taking a dual perspective on the Hard Problem of Consciousness

5.1 Introduction: What *seems* to be the problem?

Every Friday after work, three friends meet up at their local pub for a few drinks. One evening, to celebrate one of the friends' birthday, they each order a €10 whisky. They pay the €30, but since the barkeeper is in a generous mood, she gives her three regular customers €5 in return, charging only €25 for the three drinks. Each of the friends receives €1 in change, leaving €2, which they of course can't divide by three. In other words, instead of having paid €10, they each had to pay only €9, leaving them with a remaining €2. Now, three times €9 is €27. €27 + €2 = €29, but, oddly enough, not €30. What happened to that final euro?

The above puzzle is an example of what is known as a mathematical misdirection puzzle[95]. The answer to the puzzle is that we need no answer, because there really is no problem here. The idea that we need to find a missing euro is misguided. There is no missing euro, just as there is absolutely no reason to add up the different amounts. In fact, the above puzzle is more like a conjuring trick, where the audience's attention is being misdirected. In this case, the misdirection lies in making the reader assume that the question at the end is relevant, when actually, it is not.

There are people who believe that the puzzle known as the Hard Problem of Consciousness (henceforth: HPC) is in some respects similar to the scenario

[95] There is a long and rich history of such mathematical misdirection puzzles. For more details, see David Singmaster's impressive online archive: http://www.puzzlemuseum.com/singma/singma6/SOURCES/singma-sources-edn8-2004-03-19.htm#_Toc69533836.

above. The question, "How does conscious experience arise out of matter?[96]" is thought to be misguided, pretty much in the same way the question about the missing euro is. Like the latter, the puzzle of phenomenal consciousness is considered to be a pseudo-problem, or, as the early positivists called it, a *Scheinproblem*. It is felt that the idea that science will one day be able to explain how consciousness comes into existence is as ill-conceived as the idea that we will one day explain where the missing euro went to. It is simply "a cognitive illusion of sorts" (Silberstein & Chemero 2015: 186). Unsurprisingly, these people's advice is to drop the issue. Using Levine's 'explanatory gap' terminology (Levine 1983): there simply isn't an explanatory gap, because there is nothing that needs explaining.

Other people accept that the HPC might very well be a pseudo-problem, but that this is something that is itself in need of an explanation. Perhaps the puzzle about the missing euro is indeed misguided, but that doesn't mean that there no longer is *any* puzzle to solve here. After all, we might still want an explanation of the puzzling matter of how the pseudo-problem manages to present itself as a genuine problem. The HPC may be a 'cognitive illusion of sorts', but no one ever said illusions are explanatorily untaxable. Perhaps questions about missing euro's and emerging conscious experiences are misguided, but that doesn't mean that the question as to why the HPC appears as a real problem is itself equally misguided.

[96] This is one of the standard formulations of the Hard Problem of Consciousness, although it should be noted that the materialist component of this formulation – how does conscious experience *arise out of matter* – is no necessary ingredient of the HPC. Gallagher & Zahavi, for instance, define the HPC as the problem of the "very existence of subjective experience itself; it is about the very fact that objects are given to us." In its shortest version, then, the HPC is the problem of why there is experience. In addition, some authors make a further distinction between one *absolute* and two *comparative* problems. Hurley and Noë (2003) define the absolute problem as follows: "Why should neural process be 'accompanied' by any conscious experience at all?" (Hurley & Noë 2003: 132) The comparative problem comprises an intermodal, as well as an intramodal explanatory gap: "First, there's the *intermodal comparative gap*: Why does certain neural activity give rise to visual rather than auditory experience, say? Second, there's the *intramodal comparative gap*: Why does certain neural activity give rise to experience as of red, say, rather than experience as of green?" (Hurley & Noë 2003: 132) Most authors conceive of the HPC in the sense of Gallagher & Zahavi, as well as Hurley & Noë's absolute problem. Ned Block is an exception to this. On several occasions, he defines the HPC as the problem of why the brain basis of an experience is the basis of *that* experience as opposed to another, or none. (see, for instance, Block 2006) This is Hurley and Noë's formulation of the second comparative problem. In the following, we will understand the HPC in the absolute sense as defined by Hurley & Noë, Gallagher & Zahavi, Chalmers, and many others.

People who approach the HPC from this angle typically argue that, because meeting the HPC head on is something like trying to charge the horizon, we should rather focus on the philosophical assumptions that conjured it up in the first place. Just as the whole missing euro puzzle draws its strength from the mistaken assumption that we need to get to €30 by adding the different sums, so too must we find the flaw in our set of assumptions that keeps fueling the HPC. According to a prominent view, advocated by, for instance, Michael Silberstein and Anthony Chemero, that flawed assumption is *physicalism* and *ontological reductionism*:

> [I]t helps to remind ourselves what metaphysical assumptions go into generating the problem in the first place. In a nutshell, the answer is physicalism and ontological reductionism. (Silberstein & Chemero 2015: 184)

At least in this respect, these authors seem to be in agreement with David Chalmers, who famously coined the phrase 'Hard Problem of Consciousness'[97]. For Chalmers too believes that the HPC only arises against a physicalist background. An important difference, however, between Chalmers and other philosophers – to whom we will turn below – is that the former appears to accept the HPC as a genuine problem for which a genuine solution can be found. When it comes to explaining consciousness, according to Chalmers, there really is a puzzle to puzzle out. We just have been approaching it from the wrong angle. But in principle, the HPC is the name of a problem that needs to be solved, not merely dissolved.

[97] It should be noted that, as Chalmers himself acknowledges, the philosophical conundrum he calls the Hard Problem of Consciousness had already been articulated long before Chalmers came up with the terminology. In 1868, for instance, Thomas Huxley writes: "How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp." (Huxley 1868: 178) But still older mentions of the problem can be found, for instance in the work of Newton, Locke, Mill and Leibniz. In 1714, the latter famously writes: "Moreover, it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think,
feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so
that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception." (Leibniz 1714: section 17). Closer to Chalmers' day, as already mentioned, Joseph Levine referred to the Hard Problem in terms of an "explanatory gap" (Levine 1983).

This sharply contrasts with the view nowadays advocated by authors like Daniel Hutto, Erik Myin and Michael Kirchhoff. These philosophers explicitly deny that, when it comes to the HPC, there is anything to solve or explain. They write:

> The only effective way of dealing with the hard problem is to deny, from the off, "that there is a relation between the phenomenal and the physical that needs explaining" (Kirchhoff & Hutto 2016: 308; Hutto & Myin 2013: 169).

When it comes to "hard problems" or "explanatory gaps", these authors endorse a strict identity hypothesis when it comes to the relation between the phenomenal and the physical:

> We argue for adopting a strict identity thesis, one that allows for no gap between the phenomenal and the physical. (Kirchhoff & Hutto 2016: 304)

Moreover, in their book *Radicalizing Enactivism*, Hutto & Myin suggest that we

> take phenomenality to be nothing but forms of activities—perhaps only neural—that are associated with environment-involving interactions. If that is so, there are not two distinct relata—the phenomenal and the physical—standing in a relation other than identity. Lastly, come to see that such identities cannot, and need not, be explained. If so, the Hard Problem totally disappears. (Hutto & Myin 2013: 169)

These authors hold that the HPC is indeed much like our missing euro puzzle above. First, it is a pseudo-problem that requires dissolving, rather than solving. Second, this can be managed if we assume a strict identity between the physical and the phenomenal, for then it no longer makes sense to want to explain the phenomenal in terms of the physical. I agree to both points: the HPC is a pseudo-problem that needs to be dissolved, rather than solved, and second, our best option here is to adopt a strict identity between the mental and the physical. I'll have a lot more to say about this second point below, but first, I want to say something about why I think the HPC is, indeed, a pseudo-problem.

5.2 The Hard Pseudo-Problem of Consciousness

I want to approach the matter via a classic articulation of the HPC we find in a famous paper by Frank Jackson (1982). In it, Jackson develops his illustrious 'Knowledge Argument' against physicalism. The argument is supposed to show that, even if we would have a complete physical understanding of the universe, the qualitative character of experience (Thomas Nagel's 'what-it-is-like' character) would not be a part of this complete knowledge. I'll rehearse Jackson's argument below. It should be noted, however, that Jackson's argument also has its philosophical precedents, for instance in the work of C.D. Broad (1925) and Herbert Feigl (1958). Since Feigl is a central figure in our present discussion, it seems only appropriate to also present his more original version of Jackson's Knowledge Argument. In his essay *The Mental and the Physical* – to which we shall return more than once throughout this paper – we find these lines:

> We may ask, for example, what does the seeing person know that the congenitally blind person could not know. Or,[...], what could a someone know about the effects of jokes if he had no sense of humor? Could a Martian, entirely without sentiments of compassion and piety, know about what is going on during a commemoration of the armistice? For the sake of argument, we assume complete physical [...] predictability and explainability of the behavior of humans equipped with vision, a sense of humor, and sentiments of piety. The Martian could then predict all responses, including the linguistic utterances of the earthlings in the situations which involve their visual perceptions, their laughter about jokes, or their (solemn) behavior at the commemoration. But *ex hypothesi*, the Martian would be lacking completely in the sort of *imagery* and *empathy* which depends on familiarity (direct acquaintance) with the kinds of *qualia* to be imaged or empathized. (Feigl 1958: 64)

In Jackson's own thought experiment, the role of Feigl's congenitally blind person is played by Mary, the brilliant color scientist who knows everything there is to know about the physics of color and color vision, but who has never seen anything but black and white objects in her black and white room. The twist in Jackson's version is that, unlike the blind person, eventually, Mary is able to see colors when she is let out of her room. Jackson then confronts the reader with the

question: after seeing a ripe tomato for the first time, or more to the point: after experiencing redness for the first time, does Mary gain new knowledge? Does she now know something about the color red that she didn't know before?

The thought experiment is originally meant to show that it is in principle impossible to obtain a complete physical understanding of experiential consciousness. Even if we would one day have a completed physics, some important part of knowledge would still be missing, namely the knowledge of what it is like to experience something. Complete knowledge of the physics of color does not include knowledge of qualities (the redness of red, for instance). From this, it is concluded that facts about experiential consciousness can't be physical facts. In other words, there can be no such thing as complete physical knowledge of the universe that also manages to captures our qualitative experiences.

5.3 Dismantling Mary

Jackson's Knowledge Argument has proliferated into a "cottage-industry literature" (Walter 2002: 104). My own contribution to this literature consists in emphasizing that the whole problem is, indeed, a *Scheinproblem*. Jackson's thought experiment fools the reader (apparently including Jackson himself) into believing that our experiences can somehow be expected to be a matter of our theoretical knowledge. For saying that something still escapes our theoretical knowledge is suggesting that we might sensibly expect it to be part of it. The misdirection lies in the acceptance of this idea, i.e., that it is reasonable to expect there to be a connection between what we know and what we experience, such that, if we would have complete (physical or other) knowledge of our experience, we would somehow *have* the experience (for knowing what an experience is like means having, or having had the experience). In other words, the conjuring trick consists in making the reader believe that experience is on a par with theoretical knowledge. But this is simply wrong. Having an experience does not equate to having knowledge. It equates to having something we can have knowledge about.

For example, an experience of red is not knowledge of red, it is what makes it the case that there is something to have knowledge about (what the color red is like). But the experience itself is not part of the knowledge. In a sense, then, the phrase 'knowing what it is like' is misleading, for it suggests that the occurrence of the experience entails knowledge. It doesn't. Again, the occurrence of an experience is something that can become the subject of our knowledge, or what the knowledge is about, but it makes no sense to assume that experiences can themselves be part of the complete list of true (physical) propositions. Why would we ever believe that our theoretical knowledge of some subject should somehow produce that subject itself? Even if we would accept the idea of some individual possessing complete physical knowledge about something (say, color vision) as a sensible idea – which is already granting a lot, for it assumes among other things that the set of potential physical knowledge must be finite – it would still be absurd to therefore assume that this knowledge somehow produces or collapses into the known thing itself (color vision). So if we accept that Mary knows everything there is to know about color vision, why would we ever accept the very different idea that she *therefore* should *have* color vision?

Jackson's Knowledge Argument, then, is a pseudo-problem in that it is built on the misguided premise that complete theoretical knowledge about something might be expected to collapse or turn into the very thing the knowledge is knowledge of. In short: it is built on a conflation between *knowledge* and *known*. The fact that most people – including Jackson – would conclude that Mary gains new knowledge after experiencing redness for the first time, and that therefore complete physical knowledge of the universe misses an important part is misguided: Mary doesn't gain new knowledge, she gains new subjects to have knowledge about (color experiences). As for the contrary response that claims that Mary does already know what it is like to experience redness after leaving the room, this is equally misguided. Both opposing views are built on an agreement about the same misguided premise, namely, that the amount of knowledge we have of something – regardless of whether it is physical knowledge or some different kind – should somehow contribute to the actual occurrence of

that something. But this is something theoretical knowledge should *never* be expected to do, regardless of the subject matter, i.e., what the knowledge is knowledge *of*. In addition, it is also unclear why we would need the artificial and implausible set-up of Jackson's thought experiment. We don't need the 'Mary's room' scenario to bring across the idea that we can have new experiences, regardless of how much theoretical knowledge we have. Most of us know what it is like to experience red, but most of us don't know what it is like to eat snake, or to have Ebola, or to experience zero-gravity. Taking the latter as an example, accepting Jackson's argument as sensible means that we should also accept as sensible the possibility that, because someone knows everything there is to know about zero-gravity, we might expect her to experience weightlessness (or even to start floating around) and that the fact that this doesn't happen means that her knowledge about zero-gravity is incomplete. But the fact that complete knowledge of zero-gravity doesn't entail the experience of weightlessness doesn't mean that zero-gravity can therefore not be physically understood, just as the fact that, since Mary has no color experience, color experience is not susceptible to an understanding in physical terms. In general, the fact that an assumed complete physical knowledge of experience does not entail, or does not collapse into the experience itself shouldn't make us accept the idea that phenomenal consciousness evades physical explanation. This is not to say that we therefore *can* explain consciousness in physical terms. It is to say that Jackson's Knowledge Argument gives us no reason for accepting the contrary. By extension, this also means that we have no reason for accepting the terms of the debate about the HPC, as exemplified in Jackson (1982), and many other versions of the problem. The 'good reasons' we think we have are, as I've tried to argue, the result of somehow being misdirected. And, as we've seen, the central misdirection which gives rise to the idea that the existence of phenomenal experience brings with it a Hard Problem lies in accepting that our physical explanations of an entity (e.g. phenomenal experience) should somehow collapse into the very entity itself. One way of avoiding the misdirection is by assuming a strict identity relation between the phenomenal and the physical. How this is to be understood exactly will be clarified below.

5.4 Dissolving the HPC with identity? A potential tension

Prima facie, putting forward identity theory for dissolving the HPC does not sit well with the observation that the HPC arises against the background of physicalism and ontological reductionism. To the extent that identity theory qualifies as a physicalist theory – which the classic formulations of Ullin Place, Herbert Feigl and John J.C.C. Smart were certainly meant to do[98] – there is a potential tension because, as we've just seen, physicalism is being debunked as precisely the breeding ground for the HPC. At the same time, however, a physicalist theory (identity theory) is put forward as a way – perhaps even *the only* way – to adequately deal with the problem of the "very existence of subjective experience itself" (Gallagher & Zahavi 2008: 188). However, as I've also said, the tension is only *prima facie*. The problem quickly loses its pertinence once we remind ourselves of the ambiguous nature of terms like 'physicalism', 'materialism', and other related notions (e.g. 'physical'). What philosophers like Chalmers, Silberstein and Chemero have in mind when they point the finger at physicalism is the kind of ontological reductionist physicalism we find expressed in claims such as:

> All the processes in the universe, from atomic to bodily to mental, are purely physical processes involving fermions and bosons interacting with one another. Eventually, science will have to show the details of how the basic physical processes bring about us, our brain, and our behavior. But the broad outlines of how they do so are already well understood. (Rosenberg 2011: 20)

For many, Rosenberg's optimism of one day being able to solve the HPC in terms of interacting fermions and bosons is similar to that of someone who is confident that there is nothing 'pseudo' about the puzzle of the missing euro and that, in addition, we will one day be able to locate it. This is why, as we've already seen, in discussing the assumptions that generate the HPC, Silberstein and Chemero don't simply speak of physicalism, but of physicalism understood as "ontological reductionism." (Silberstein & Chemero 2015: 184) Crucially, however, one can be

[98] See, for instance, Smart 2017: "In taking the identity theory (in its various forms) as a species of physicalism, I should say that this is an ontological, not a translational physicalism."

an identity theorist without therefore having to pledge allegiance to the gratuitous credo: 'the physical facts fix all the facts'. What I want to argue, then, is that identity theory does not necessarily equate to ontological reduction. This, however, is still a wide-spread assumption. Yet, when an identity theorist like Herbert Feigl – arguably the most overlooked classic identity theorist – presented his theory as in line with physicalism, a commitment to ontological reductionism was *not* what he had in mind when he suggested the idea that mental processes are strictly identical with physical processes. What I want to make clear here is that, despite courant readings, identity should not be conceived of as entailing reduction and that, in the end, a properly understood identity theory is still the best way to tackle the HPC or the problem of experience.

5.5 Identity and reduction

Within the literature, there is a strong tendency to lump identity theory together with reductionism. Admittedly, the way to a reductionist reading of identity appears to have been paved by classic identity theorists themselves, especially by Smart. Consider these citations:

> I shall be concerned to put man in his place by defending the view that he *is nothing more than* a complicated physical mechanism. (Smart 1963: 15, m.e.)

> I wish to argue for the view that conscious experiences are *simply* brain processes. (Smart 1963: 88, m.e.)

We see these identity claims reappearing in more recent writings, for instance by committed materialist thinker Valerie Hardcastle:

> I say that if we are materialists, then we have to believe that consciousness is something physical. Presumably it is something in the brain…the mind *is nothing more* than activity in the brain. (Hardcastle 1996: 8-9, m.e.)

The reductionist nature of these assertions is unmistakable. The problem, however, is that claims like 'consciousness is something physical' or 'mind is nothing more than brain activity' are ambiguous. Understood in a reductionist

way, what could it even mean to say that consciousness is simply a brain process, or that consciousness is something physical? Does it make sense at all to think that, every time I have a conscious experience, what I'm really having is not a conscious experience, but 'nothing but' a brain process of some kind? For instance, that we would say that we are not really feeling pain, but that there's nothing but a neural event *x* going on in our brain. And what motivates this, what Donald Davidson calls, 'nothing-but reflex' (Davidson 1980: 214)? Furthermore, if conscious experiences are really strictly identical with brain processes, as the classic identity theorist claims, why not hold that these brain processes are 'nothing but' conscious experiences, rather than the other way around? After all, as Davidson also points out:

> [I]f some mental events are physical events, this makes them no more physical than mental. Identity is a symmetrical relation. (Davidson 1987: 453)[99]

The above formulations by Smart and Hardcastle are starting to look a little too much like the radical physicalist's claim that reality, including consciousness, is 'nothing more' than interactions of fermions and bosons. So if these formulations are truly representative of identity theory, then identity theory indeed appears to be cause, rather than cure for the HPC.

Fortunately, identity theory does not stand or fall by Smart's reductionist formulations. As already indicated, I do not think that strict identity statements entail reductionism. To see how mind-body identity claims do not collapse into ontological reductionism, it is important to highlight that an identity statement like 'A just is B' is, like some of the above formulations, ambiguous in that it allows for two interpretations. It may mean that, whatever we thought to be B (something 'mental') is actually something else, namely A (something 'physical'). In this sense, the mental collapses into the physical, but not the other way around. Call this the *ontological reductionist* reading. On another reading, however, the statement 'A just is B' may mean that, because of A and B's identity,

[99] In this regard, it should be stressed that Davidson's anomalous monism is not to be understood as a form of physicalism or materialism, be it reductive or non-reductive. Davidson is very explicit about this: "Anomalous Monism is not a form of physicalism or materialism." (Davidson 1995: 75).

the terms do not pick out a real distinction, but one and the same thing which is, of course, irreducible to either A or B. To see the crucial difference better, consider an analogy. Saying that the Evening Star *just is* the Morning Star may mean that there really isn't an Evening Star, but only a Morning Star. But it may also mean that both names pick out just one entity. In this second sense, the idea of one entity being ontologically reducible to the other simply doesn't apply, because there is no real distinction to begin with. It is in this second, non-reductive sense that we want to understand and defend identity theory. Strict identity, as we understand it, does not entail reduction to one of the relata of the identity relation. Logically formalized:

$$(A = B) \neq (A \wedge \sim B) \vee (B \wedge \sim A)$$

Smart rightly claims that 'you can't correlate something with itself' (Smart 1959: 142). Similarly, it is equally clear that you can't reduce something to itself.

5.6 Identity Theory and Neutral Monism/Dual Aspect Theory

John Heil observes that the idea of there being no real distinction between the mental and the physical (endorsed by, for instance, Davidson) goes at least back to Spinoza:

> The mental–material distinction is, as Spinoza[100] and Donald Davidson contend, a distinction of conception only, not a real distinction, not a distinction in reality. (Heil 2013: 185)

Identity theory accepts that the mental-material distinction does not pick out a real, mind independent distinction. The distinction is a result of our world-structuring activities, not something independent of these activities. However, to the extent that the idea of there not being a real distinction is understood as

[100] It might not be entirely accurate to attribute this non-realism with regard to the mental/physical distinction to Spinoza. For him, the mental and the physical are understood as two of an infinite set of 'attributes'. I doubt that Spinoza would have agreed to attribute only a conceptual reality to this set.

advocating neutral monism and/or[101] dual-aspect theory, our position differs in that we do not believe in the existence of a neutral substrate, conceived as neither mental, nor physical. The main reason for our disagreement with this view, defended at present by philosophers like Silberstein and Chemero, as well as with dual-aspect theory is that the notions of the physical and the mental are here still thought of as either

1. possible objective characterizations or descriptions of a substrate, albeit negative characterizations. (neutral monism), or

2. as positive objective characterizations or descriptions of two different aspects or properties of a neutral substrate (dual- aspect theory)

The worry is that both this kind of neutral monism and double-aspect theory reiterate a misguided logic in which 'physical' and 'mental' are essentially descriptive terms that can be used to objectively describe an entity's nature (i.e., as it really is, independent of our relation to the entity), similar to the way ordinary adjectives are used to describe or characterize an object. I think this is problematic for a number of reasons.

First, this logic still leaves us with the idea of some mysterious object or substrate to which both descriptions 'mental' and 'physical' apply, either as a negative qualification of this substrate's nature (for neutral monism thinks of it as *neither* mental *nor* physical), or as a positive qualification of *two aspects* of this neutral underlying reality (dual-aspect theory). However, I do not agree with the 'two-sides-of-the-same-coin analogy' (Silberstein and Chemero 2015: 191), for there is no coin. The idea of, what Feigl calls, a 'neutral third' (not in the sense of there being three neutral substances, but of there being three possible characterizations, namely physical, mental and neutral) raises unnecessary questions and does not sit well with the principle of parsimony. I agree with Feigl when he writes:

[101] There is still debate about whether dual-aspect theory is to be counted as a form of neutral monism, hence the "and/or".

> If the neutral third is conceived as unknown, then it can be excluded by the principle of parsimony which is an essential ingredient of the normal hypothetico-deductive method of theory construction. If it is defined as in principle unknowable, then it must be repudiated as factually meaningless on even the most liberally interpreted empiricist criterion of significance. (Feigl 1958: 82-83)[102]

Second, and more straightforwardly, I think it is simply a mistake to assume that terms like 'mental', 'psychical', 'material', 'physical' or 'neutral' pick out possible objective characteristics of an entity *at all*. I agree that these adjectives can be properly used to characterize our characterizations, in the sense that we can distinguish between a mental or a physical discourse, but I disagree with the idea that the terms pick out real properties or qualities of an entity which it has in and of itself (e.g., in the traditional sense in which a stone is said to have physical, but no mental properties, a mammal is said to have both physical and mental properties, and a soul is said to have only mental properties). The point I'm trying to make can perhaps be best summarized by saying that the distinction between the adjectives mental/physical presents us with a false dichotomy. An entity has no physical properties, as opposed to – and therefore, on a par with – mental properties. Rather, attributing physical or mental properties to an entity should be understood as relative to two different perspectives we can have on an entity. But it is not that there is a perspective-independent object which, from one perspective, shows us its mental properties, whereas from the other perspective its physical properties. Indeed, from the intersubjective perspective, an object appears as an object with physical properties. But from the perspective of experience, we do not find an object with a different kind of properties (mental). Rather, the so-called mental properties apply to our experiences of, what from the intersubjective perspective appears as, objects. 'Mental', therefore, does not

[102] We find a similar rejection of the usefulness of this idea of an underlying neutral reality in Nelson Goodman's *Ways of Worldmaking* (1978): "So long as contrasting right versions not all reducible to one are countenanced, unity is to be sought not in an ambivalent or neutral something beneath these versions but in an overall organization embracing them." (Goodman 1978: 5) And further: "But what is it that is so organized? When we strip off as layers of convention all differences among ways of describing it, what is left? The onion is peeled down to its empty core." (Goodman 1978: 118)

apply to objects or their properties, but to our experience of these objects and their properties. Property dualism, as we find it nowadays for instance in Chalmers, is therefore mistaken. It is based on the mistaken assumption that mental properties can be brought on a par with physical properties. On closer inspection, however, we see that it is precisely this mistaken assumption which underlies, and in a sense unites, dualist and monist approaches alike. Whereas the dualist believes the mental and the physical to pick out two distinct realities, the monist accepts the terms of the dualist, but claims

1. that there is only physical reality (ontological reductionism or eliminativism about the mental), or
2. that there is only mental reality (ontological reductionism or eliminativism about the physical), or 3. that there is only one non-mental and non-physical reality, but which has aspects characterizable as mental or physical (neutral monism/dual-aspect theory).

All these positions accept and reinforce the terms of the debate: 'mental' and 'physical' are notions that can be used to describe the nature of an entity or that entity's properties (even when that entity refers to reality itself), regardless of our perspectival relation to the entity.

Another worry with neutral monism is that, despite the stipulated neutrality of the underlying reality, all neutral monists seem to believe that we can still somehow *know* this reality, though perhaps only in a very limited sense. This comes out in the specific names each neutral monist uses to designate this neutral reality. William James spoke of 'pure experience', Ernst Mach used the term 'sensations', Bertrand Russel coined the term 'sensibilia', and more recently, authors like William Seager, Michael Silberstein & Anthony Chemero prefer the term 'presence'. The problem is not so much that these neutral monists "all choose to use terms that are decidedly non-neutral in tone." (Seager 2013, as quoted in Silberstein & Chemero 2015: 186) Terminology can easily be fixed. The problem, rather, is that these interpretations of neutral monism all seem to

believe that intelligibly referring to a reality behind or beneath perspectival reality, from a Nagelian view from nowhere, is within the limits of our cognitive capacities. Take, for instance, Seager's notion of presence, which is adopted by Silberstein and Chemero. Apparently, we seem to know quite a lot about this neutral reality. Seager tells us that

> [p]resence is what constitutes the 'what it is like' of conscious experience….it forms the bedrock of reality…and may yet help to understand reality and our place within it. (Seager 2013, as quoted in Silberstein & Chemero 2015: 192)

And Silberstein & Chemero inform us:

> Presence can be thought of as temporality or 'nowness' itself; there is nothing phenomenologically more basic than the nowness or presence of experience. Perceiver (subject) and perceived (object) are *co-dependent aspects of presence*…. Moreover, presence is not the product of perception and action, as the sensorimotor view sometimes suggests. Presence is the field of experience that gets modulated and modified by perception and action. (Silberstein & Chemero 2015: 193; m.e.)

It is unclear how these speculative claims are supposed to be interpreted. Furthermore, rather than solving, they only generate new conceptual problems. What is "'nowness' itself", for instance? And in what sense are we to understand the claim that presence can have "co-dependent aspects"? I think that much of the metaphysical problems associated with dual-aspect theory, as well as forms of neutral monism, can be avoided by shifting the focal point from 'known reality' to the 'knower', or from 'aspects of reality' to 'perspectives on reality'. Rather than assuming the existence of different kinds of realities (mental or physical) that, from a certain perspective, show up, or precisely fail to show up (neutral), we should take one further step back and redirect our attention to our perspective-taking practices themselves. To avoid the abovementioned problems with which neutral monism and dual-aspect theory are confronted, we might turn to the classic identity-theoretical approach suggested by Feigl. Instead of dual aspects,

Feigl thinks it is "wiser to speak instead of twofold access or double knowledge." (Feigl 1958: 80) In a similar vein, Peter Strawson speaks of our capacity to occupy "alternative standpoints", namely the "scientific-objective standpoint" and the "human-perceptual-and-moral standpoint" (Strawson 1985: 55). Indeed, the identity theoretical view I want to defend may very well be dubbed *dual-perspective* or *dual-standpoint theory*. It claims that states or events experienced from within the subjective perspective are identical with the referents of (certain) terms of the language of physics, i.e., as 'seen' from an objective, or rather, intersubjective perspective.

This dual-perspective view is also being endorsed by philosopher Erik Myin, who aptly formulates the position when he suggests that we

> understand the philosophical difficulties surrounding the relation of perceptual experience to the physical as deriving from a difference in perspectives, rather than an ontological difference — it construes what dualists conceptualize in ontological terms as a distinction between perspectives. (Myin 2016: 85)

However, on Myin's account, the physical perspective includes, *but is not restricted to*, what can be captured in neurophysiological terms. It is in this latter respect that this version of the dual-perspective view somewhat digresses from Feigl and other identity theorists. Most identity theories – both classic and contemporary versions – are neurocentric. However, a more promising approach is to try to understand phenomenality from the physical perspective by 'going wide'[103]. In the next section, I'll explain how this is to be understood by contrasting it with the idea that the relation between the phenomenal and the physical can be understood as a supervenience relation.

5.7 Identity and supervenience

In the concluding chapter of his *Action in Perception*, Alva Noë writes:

[103] The phrasing 'going wide' is borrowed from Hutto & Myin. See Hutto & Myin 2013: 151, 157, 158.

> Most cognitive scientists hold that for every experience there is a neural structure or substrate whose activation is sufficient for the experience. On this way of thinking, experiences are internal biological processes, comparable to digestion and respiration; they happen inside us. Philosophers sometimes put the basic idea like this: Neural duplicates are necessarily experiential duplicates, for experience *supervenes* on states of the brain. (Noë 2004: 209)

Throughout the chapter, Noë argues against the kind of neurocentric internalism that has been dominating cognitive science since its inception. Instead, he wants to put forward, what he calls, "enactive externalism" (Noë 2004: 221). Enactive externalism wants to take serious the idea that "experience might *depend* constitutively on physical substrates that are not inside the head (e.g., on dynamic patterns of interaction among neural processes, the body, and the environment)." (Noë 2004: 210; m.e.) This dependency relation is cast in terms of supervenience: "…experience might supervene not on the brain, but rather on brain-animal-world systems." (Noë 2004: 218), where these systems are conceived of as the causal basis for experience:

> [E]xperience might not be in the head. Whether it is depends on whether, as a matter of fact, the *causal basis* of experience depends on ongoing, causal interaction among brain, body, and environment." (Noë 2004: 219; m.e.)

Put otherwise, Noë has no quarrel with the way the nature of the phenomenal-physical relation is conceived of. This, he accepts, is a causal supervenience relation. Rather, his disagreement concerns the supervenience *basis*. Instead of restricting phenomenality's causal basis to what goes on in brains, Noë proposes a wide supervenience basis. Experience doesn't supervene on neural events alone, but on the embodied dynamical interactions of the situated organism, which include neural activity.

In light of present purposes, we shouldn't be too concerned with the exact empirical nature of Noë's 'going wide' proposal. At this point, I also won't be considering the question how, if at all, enactive externalism can help solve or dissolve the HPC (but see below; see also Prinz 2006: 17). Here, I particularly want to draw attention to the way the relation between experience and the

relevant physical structures (i.e., brain-body-environment relations) is being grasped. For despite his dismissal of mainstream cognitive science's neurocentric and internalist assumptions, by invoking a supervenience relation, Noë buys into a different wide-spread assumption of mainstream cognitive science, which might be equally problematic: the assumption that the experiential level *causally supervenes* on an underlying physical substrate, flanked by the assumption that this idea is somehow explanatorily useful. Unfortunately, it is far from clear what the supervenience relation is supposed to be, nor how we should understand its explanatory relevance. These two critical observations can also be found in the work of philosopher Jaap van Brakel, to which I shall return in the next section.

5.8 Supervenience: ambiguity and vagueness

Although older sources can be cited, the term 'supervenience' is usually traced back to the work of Richard Hare and Donald Davidson.[104] Within philosophy of mind, however, it probably gained most familiarity through the work of Jaegwon Kim (see, especially, Kim 1993). The common thread throughout the term's various uses is that of a relation between a base level *B* and a higher level *S*, where *S* in some sense depends on the base level *B* such that any change to *B* entails change to *S*. The central question as to how this dependency is to be understood remains a matter of debate. However, perhaps a more pertinent question is what the supervenience relation is supposed to be doing in the first place. Jaap van Brakel characterizes the notion's role succinctly. Supervenience is supposed to relate "different levels of analysis or discourse" (van Brakel 1996: 253). It is a way of addressing the problem of, what he calls, 'interdiscourse relations'. To be sure, relating a *mental* discourse to a *physical* discourse is but one of the notion's applications, and van Brakel gives plenty examples that show how the notion of supervenience and its synonyms[105] are invoked to solve problems of

[104] See Hare (1952) and Davidson's *Mental Events* (1970, reprinted in Davidson 1980). However, as van Brakel (1999) points out, a significantly earlier source can be found in a 1926 *The Journal of Philosophy* paper by Stephen C. Pepper. See Pepper, S.C. (1926).

[105] There are many other terms besides 'supervenience' to give expression to this relation: "For example one might say about the relation between an S- and B-discourse or domain that *B fixes S;* that S

interdiscourse relations, not only within philosophy of mind, but in virtually all of the different science domains. For instance, in physics, macrophysical properties (e.g., brittleness, having a temperature T) are said to supervene on microphysical ones (respectively, molecular structure and kinetic energy of constituent molecules).Within biology, we find the idea of genes supervening on DNA sequences, or species selection supervening on organismic selection. But we also find examples of supervenience in the 'soft' sciences, for instance, when the social is said to supervene on the actions of individuals, or when the aesthetic quality of a painting is said to (partially) supervene on physical properties.[106] Yet, no matter in which area the notion of supervenience is being called upon, there's no doubt as to the job it is supposed to be performing in each case, namely gluing together different discourses. At the same time, however, it remains unclear *how* supervenience is supposed to be doing this. As van Brakel points out, the problem with the idea of levels supervening on, or underlying other levels, is that it is inherently ambiguous and vague:

> It is often unclear whether saying that B underlies S means that *B causes S,* or that *S is the same as B* (but at a different level of description), or that S should be understood or explained in terms of B, or whether it is left completely vague or indeterminate what the relation is between the S- and B-discourse. (van Brakel 1996: 258)[107]

Noë's appeal to supervenience is *prima facie* characterized by the very same ambiguity. When it is claimed that dynamical brain-body-environment structures are the causal basis for experience, with the latter depending on the former, this can be taken to mean at least two things. On the one hand, 'causal' might be used here to specify only the relations holding between the physical substrate, while leaving the dependency relation (i.e., supervenience) indeterminate. On the other hand, 'causal' might also be interpreted as applying to the dependency

---

is *brought about by* or *realised in B* phenomena; that *S* properties are possessed *in virtue of B* properties; that *S* is *grounded in B;* that *B underpins S;* that *S* is *determined by* or *dependent on B;* that *S boils down to* or *comes to nothing more* than B; that *B controls S;* and even that the *constitution of S* is a function of the *constitution of B.*" (van Brakel 1996: 258; emphases in original).

[106] All these examples are mentioned in van Brakel 1996: 262-264.

[107] According to van Brakel, "the vagueness exemplifies a kind of covering up in order to avoid properly addressing the question of interdiscourse relations." (van Brakel 1996: 258)

relation between the physical supervenience basis B and the supervening experiential level S. On this second reading, supervenience is incompatible with identity, understood as "S being the same as B, but at a different level of description". At times, Noë clearly endorses this second reading: "…experience depends causally on physical processes involving the extra-neural world." (Noë 2004: 223) Of course, defending an identity relation implies a refutation of this view. If experience is strictly identical with physical events (i.e., events that can be in principle intersubjectively observed), then it makes no sense to think in terms of one causing the other. Furthermore, assuming a causal relation between the physical and the experiential does not help us with the HPC, it merely fans the flames. Assuming that experience is somehow caused by physical events begs the question of *how* phenomenal experience fits into the causal chain of a physical universe, which just is a formulation of the HPC.

At the same time, however, despite Noë's explicit endorsement of a causal supervenience relation, we are being told that "[e]xperience…is something we do; it is a temporally extended process of skillful probing." (Noë 2004: 216) Admittedly, I am not sure how we are to understand these identity claims. Exchanging 'experience' here for, e.g. 'pain-experience' would entail that 'pain is something we do'. It is insufficiently clear how such a claim should be interpreted and how it relates to Noë's postulation of a causal supervenience relation. What is clear, however, is that, to the extent that the mentioned "temporally extended process of skillful probing" is to be identified with the dynamical brain-body-environment interactions, the claim would be incompatible with the idea of these physical structures causing experience. If temporally extended processes of skillful probing *just is* experience, the former cannot be the latter's cause.

In any case, on Noë's account, wide supervenience is to be understood in the sense in which experience causally depends on a "complex network involving the brain, body, and the environment" (Noë 2004: 214). Therefore, experience cannot be identified with these structures (under a different description). I want to suggest, however, that this latter possibility is actually closer to the truth. For although Noë's account remains stuck with the idea of the phenomenal being

caused by the physical, there is a valuable element in his proposal of looking beyond the brain. Indeed, experience involves more than neural activity, but the relation is not supervenience, it is identity. The central claim I want to defend, then, is that experience is strictly identical with what, from an objective or intersubjective or physical perspective, shows up as situated and embodied organismic activity, including neural activity. This requires some clarification.

5.9 Why mind the gap?

Before going further into the identity theoretical account I want to advocate, it is useful to rehearse why the HPC cannot, and does not need to be solved. If the HPC is the problem of giving a physical explanation of experience, then the problem is indeed intractable. It is, however, important that we are clear about the reason for this intractability. One way of putting it would be to argue as follows: Physical explanations require physical phenomena. In terms of the distinction between *explanandum* and *explanans*, a physical *explanans* only applies to a physical *explanandum*. For something to qualify as a genuine *physical* explanation of a phenomenon, the phenomenon under investigation must also be grasped in physical terms. And indeed, we have become rather proficient in subsuming phenomena under the general header 'physical phenomenon'. For instance, saying that temperature is the energy of the different motions of particles[108] is not so much an explanation of temperature, but rather a physical redescription allowing the phenomenon to be introduced into a greater causal explanatory framework. For many, the idea that we can do something similar with mental phenomena is nothing short of a category mistake. They argue that mental phenomena are non-physical by definition, so they are by definition excluded from potential physical redescriptions. I do not agree with this a priori line of reasoning. Making scientific progress dependent on pre-existing definitions is never a good idea. But more importantly, I simply don't think that

[108] I am aware that, understood as a type-type identity claim, the timeworn example that 'temperature is kinetic molecular energy' is imprecise and outdated. Here, however, it is not meant to be read as implying an endorsement of this (false) type-type identity claim.

'physical phenomenon', as opposed to 'mental phenomenon', picks out a real category at all. As already emphasized above, I think it is a mistake to conceive of the terms 'physical' and 'mental' as objectively characterizing a phenomenon. Rather, it characterizes our descriptions of, and perspectives on a phenomenon. To better explain what is meant here, it is helpful to contrast this view with Herbert Feigl's approach to the notions 'physical' and 'mental'. Feigl defines 'physical' as

> the sort of objects or processes which can be described (and possibly explained or predicted) in the concepts of a language with an intersubjective observation basis. This language or conceptual system is—in our sort of world—characterized by its spatio-temporal-causal structure. (Feigl 1958: 54)

Feigl is right when he insists on intersubjective observability as an essential feature of our understanding of what it is to be a 'physical phenomenon', as opposed to a 'mental phenomenon'. However, I think Feigl's definition is nevertheless misguided in its use of the adjective 'physical' to refer to a "sort of objects or processes". As long as we keep conceiving of the world as consisting of two sorts of entities, i.e., physical and mental entities, we'll keep paying homage to the exact same picture whose frame we should be trying to break. I therefore want to suggest a somewhat different definition of 'physical', which still retains most of Feigl's insights:

> 'Physical' refers to the sort of descriptions or conceptions in a language with an intersubjective observation basis. This language or conceptual system is—in our sort of world—characterized by its spatio-temporal-causal structure.

So why is the idea that we can give a physical redescription – and perhaps explanation – of a phenomenon like toothache by many held to be impossible? I believe that the reluctance of accepting potential physical redescriptions of certain phenomena, including 'mental phenomena', is first of all a psychological matter. It derives from the fact that we simply cannot imagine or conceive what it would be like to give a physical redescription of the qualitative nature of our experiences, *such that we can still recognize the essentially experiential nature of*

the experiential phenomenon in the redescription. Giving a physical account of, say, a toothache can indeed never capture its experiential or qualitative character – and neither should we expect it to, as our discussion of Jackson's Knowledge Argument should have made clear – so whatever else it might be describing (a neuro-physiological event, for instance), we feel it can therefore never be a description of a *toothache qua toothache*. And the same goes for the 'wetness' of water, or the 'redness' of a color, or the 'roughness' of a surface, and so on. However, this way of thinking leads us straight back to the HPC. Indeed, we seem confronted with Levine's explanatory gap between phenomena that can be accounted for – in the sense of being describable, as well as explainable – in physical terms, and phenomena which by their very nature manage to elude these accounts. And again, the way out of the (pseudo-)problem is to realize that the gap is not the result of the 'essential nature' of the phenomena, but of the 'essential nature' of our ways of looking at them. We should come to realize that it is simply impossible to look at a phenomenon *qua* physically describable/explainable and, *at one and the same time*, as evaluable in a qualitative/mental vocabulary. To use Wilfrid Sellars' well-known analogy[109], trying to fuse these different perspectives into one stereoscopic image *while at the same time retaining the particular character of both perspectives* is not only impossible, it is also unnecessary. It is impossible, simply because you can't at once have two different perspectives in one single perspective. Believing that this is possible is like believing that we can mend something by breaking it, or that we can focus on a background, while at the same time attending to the foreground. But, as said, fusing both perspectives is also unnecessary once we accept a strict identity between what these different perspectives are perspectives on. For on this assumption, there no longer is a gap that needs to be bridged. Again, the gap only exists as a result of two conflictual compulsions: on the one hand, the idea that experiential phenomena *must* be explained by, and are only explainable within the physicalist framework; on the other hand, the feeling that, for something to count as a satisfying explanation, we need to be able to recognize something of the essential qualitative nature of the experience in the physical

[109] Sellars 1963: 4.

explanation. Exposing the HPC for the pseudo-problem it is requires that we stop acknowledging (the combination of) these requirements as legitimate.

5.10 Materialist and panpsychist responses

Viewed in this light, we see that, despite their deep differences, both materialist and panpsychist responses to the problem of experience have this much in common that they are both tailored fit to accommodate these requirements. Starting with the panpsychists, they lay down the first requirement of reducing experience to purely physical entities. Instead, their position places full emphasis on the second requirement, i.e., explaining how the phenomenality of experience can still be recognizable as a fundamental part of a physical universe. Their 'solution' is that the physical universe is in some sense intrinsically mental *by default*, or more precisely, that certain micro-entities inhering in all particular entities are. In light of the above, it should be clear why I fully reject this idea. Again, I think it is a mistake to understand the terms 'mental' and 'physical' as descriptive notions that can be used to describe properties of entities, let alone properties of elementary particles. Panpsychism seems to be doing just that:

> Panpsychism entails that at least some kinds of micro-level entities have mentality, and that instances of those kinds are found in all things throughout the material universe.[110]

In addition to the various existing objections against panpsychism, then, I think it is simply misguided to conceive of 'mentality' as something that can be had by, and distributed among different particles so that these 'mental atoms' become somehow combinable into higher forms of mentality, like conscious human experience. 'Mentality' is not something an object can have, in addition to its physical properties'. Furthermore, panpsychism explains nothing. It merely relocates the problem. Transferring a predicate (mental) that we don't really understand to the level of fundamental particles only makes those particles even more inconceivable than they already are.

[110] See Goff, P., Seager, W. and Allen-Hermanson, S. (2017).

The materialist, on the other hand, accepts the first condition of explaining phenomenality in a physicalist vocabulary. He or she can acknowledge that such an explanation is perhaps unavailable within today's physicalist framework, but who's to say what we'll discover tomorrow? And as we've seen, if we are to believe Alex Rosenberg, "the broad outlines" of a physical explanation "are already well understood". This position is seen by many as hopeless on grounds that it fails to take into account the fundamentally *sui generis* nature of experience. The point of Frank Jackson's Knowledge Argument is precisely that, even if we had a completed physics of color and color vision, we still wouldn't be able to account for how physical processes can explain the specific qualitative nature of a color: the 'what it's like' to see red would still elude us in that it would never 'show up' in the physical explanation. It doesn't matter what physical causal mechanism we discover to be underlying the generation of phenomenal consciousness (brain oscillations in the 40–70 Hz range, quantum processes in neurons' microtubules, reentrant signaling between cortical maps, and so on[111]). In other words, the second requirement of somehow recognizably capturing the qualitative nature of experience in the physical descriptions would remain unsatisfied.

My view on the matter is as follows. I agree that we will never be able to account in physical terms for the phenomenal nature of experience, if this requires that anyone having such an account (Mary, Martians or, simply, omniscient beings) may be expected to therefore have the experience as well. However, as I've already made clear above, if this is the reason for dismissing a physical explanation of experience, we should reconsider it. But I agree that we will never be able to give a causal account of experience, so that we could say that, because of such and such causal event, this is what it's like to see red. The reason is *not*, however, that experience is something that is insusceptible to our physical descriptions, the reason is that the 'what it is like character' of an experience is

[111] These are three of the more popular suggestions. For the brain oscillation hypothesis, see Crick and Koch (1990), Llinas and Ribary (1993), Singer (1993), and Singer and Gray (1995). For the quantum theoretical approach, see Penrose (1994) and Hameroff (1994). For reentrant signals, see Edelman (1989). See also Chalmers 1996 for more hypotheses about the causal structures underlying consciousness.

not *reducible* to, but *identical* to events that can be described in a physical discourse. This is why, above, I've stressed the crucial, yet easily overlooked difference between identity and reduction. The identity theory I endorse holds that, whatever we identify *as* an experiential phenomenon can in principle also be identified *as* a physical phenomenon because both can be identified *with* each other. Put differently, what is accessible via subjective experience is also in principle accessible via the (scientific) intersubjective approach. We can give a physical description/explanation of an experience *qua* physical event (i.e., *qua* intersubjectively observable and confirmable) and, vice versa, we can give a phenomenal description/explanation[112] of a physical event *qua* phenomenal event (i.e., *qua* subjectively experienced). But the idea that the occurrence of one can provide a causal explanation for the occurrence of the other makes no sense because they are strictly identical. There simply is no causal connection between two things that are seemingly different, but actually one and the same entity. So to the sceptic of physicalism, we should say: experience can be, and actually *is* being investigated from a physical perspective. And perhaps it is our best way of theoretically understanding experience. However, *pace* the ontological reductionist, the idea that we will one day be able to give a causal story of how physical phenomena cause experience *qua* experience should be put away as both impossible and unnecessary.

5.11 Lived and objective perspectives

As already indicated several times, the identity theoretical proposal I endorse is largely indebted to classic identity theorists, especially Feigl. But it is also different from these classic accounts in that it is not neurocentric. Phenomenal experience is not identical with brain states *simpliciter*. On the dual perspective account I am advocating, what is identified from the phenomenal perspective as an experience of, say, red, is to be understood as identical with what from the

[112] As an example of a phenomenal explanation, I'm thinking here of very simple examples like explaining why someone's ears hurt in terms of the loud noise she heard. The pain experience is explained by the experience of loudness.

objective/intersubjective perspective can be identified as a form of situated embodied activity. This activity includes, but is not restricted to, brain activity.

This 'going wide' proposal can in some form already be found in work by Silberstein and Chemero (see below), and it is also at the forefront of recent work by Myin. As the latter puts it:

> [T]he original identity theory as formulated in Place (1956) or Smart (1959) was narrowly brain-based, while the analysis given here emphasizes the role of bodily doings, and thus broader, body- and possibly environment-spanning identities. (Myin 2016: 88)

Like Noë, Myin (2016) also argues for a 'going wide' strategy, not in terms of supervenience, however, but in terms of identity. As we've seen in an earlier quote, Myin explicitly endorses the idea of the schism between the phenomenal and the physical as 'deriving from a difference in perspectives, rather than an ontological difference.' (Myin 2016: 85) To better show what this means, Myin instructively relates the 'difference in perspectives' to Merleau-Ponty's cardinal distinction between the lived body (*corps vécu*) on the one hand, and the objective body (*corps objet*) on the other. From the perspective of the lived body – which is, so to speak, our default setting – we encounter things in the world first of all as non-living objects from the perspective of our own lived body. But we encounter them experientially, which is to say, we encounter them through our lived bodily *activity*. When I want to pick a rose, and get pricked by one of its thorns, the sharpness of the thorns is something I encounter in virtue of my active bodily (including, in this case, my hands and fingers grabbing the flower) engagement with the object. But in these bodily engagements, our body is not itself one more object amongst objects, or, as Merleau-Ponty puts it, 'une chose entre les choses'[113]. From the perspective of the lived body, we do not relate to our

[113] It is worth quoting Merleau-Ponty at length here, for the passage including the phrase 'chose entre les choses' also contains a summary of his interpretation of the mind-body problem: "Abordons la question du rapport de l'homme et son entourage naturel ou social. Il y a là-dessus deux vues classiques. L'une consiste à traiter l'homme comme le résultat des influences physiques, physiologiques et sociologiques qui le détermineraient du dehors et feraient de lui *une chose entre les choses*. L'autre consiste à reconnaître dans l'homme, en tant qu'il est esprit et construit la représentation des causes mêmes qui sont censées agir sur lui, une liberté acosmique. D'un

body itself as an object 'out there'. Rather, we *are* this lived body. This being said, however, human beings also have this remarkable capacity of relating to the body in the same objective sense as we relate to ordinary, non-living things. It is possible for us to switch to the objective mode, also when it comes to our own bodies. We can, in other words, take up two different perspectives on our bodies and the bodies of others. Crucially, however, both perspectives can never be taken up at the same time. To illustrate this crucial point, Myin returns to Merleau-Ponty's own example[114] of one person's hand touching the other:

> One of the hands is exploring the other as object. Though a measure of ambiguity applies to both hands, the one that is touching and exploring exemplifies the lived pole, while the other hand exemplifies the objective pole. …Crucially, the same hand can't be fully touching and touched: when it switches to the touched mode, it is no longer touching; *it can't be fully lived and experienced as objective at the same time*. (Myin 2016: 84, m.e.)

Ultimately, then, the hard problem of consciousness, or the so-called explanatory gap, or simply, the perennial mind-body problem, all seem to derive from the same source. The felt schism is the seemingly inevitable by-product of this specific capacity of relating to the world from two different perspectives, together with our inability to unite these perspectives. As Myin puts it:

> From the point of view laid out here, the problem of the subjective–objective 'gap', or the abyss between the phenomenal and the physical, becomes a consequence of the fact that one can never fully take the lived and the objective perspectives at once — just like one hand cannot be both touching and touched, or like how one cannot completely 'step outside one's body' to consider it as the object of reflection. The gap becomes a fact of the human condition, without creating an ontological schism. (Myin 2016: 86)

---

côté l'homme est une partie du monde, de l'autre il est une conscience constituante du monde. Aucune de ces deux vues n'est satisfaisante." (Merleau-Ponty 1945/1948: 142, m.e.) In English translation: "The question is that of man's relationship to his natural or social surroundings. There are two classical views : one treats man as the result of the physical, physiological, and sociological influences which shape him from the outside and make him *one thing among many*; the other consists of recognizing an a-cosmic freedom in him, insofar as he is spirit and represents to himself the very causes which supposedly act upon him. On the one hand, man is a part of the world; on the other, he is the constituting consciousness of the world. Neither view is satisfactory." (Merleau-Ponty 1964: 71-72, m.e.)

[114] See Merleau-Ponty 1945: 81, 126.

The misdirection, then, lies in believing that the lived and the objective perspective *must* be united if we want to fully account for the nature of phenomenal consciousness. But, as Peter Strawson once put it, this is "to attempt a unified story where none is to be had." (Strawson 1985: 65) Indeed, what is not separated in reality does not need to be put together. All the identity theorist needs to do is elucidate as much as possible the nature of the different perspectives. And, as Merleau-Ponty helps us realize, when we take up the objective perspective to our own active experiential lives, or better: when we take up the objective perspective to our lived perspective itself, there is more than brain-activity alone that deserves our attention. Our experiential perspective is not accounted for my mere neural activity. It is also embodied in the sense that our bodies are constitutively involved in our experiential lives. We are, after all, lived *bodies*, not merely lived brains. So, to relate all this to the scientific study of experience, if we agree that scientifically investigating phenomenal consciousness requires relating to it from the objective perspective, we have no reason to assume we should restrict our attention to neural processes or brain states. In this regard, the empirical work done by, for instance, sensorimotor contingency theorists like Susan Hurley, Alva Noë and Kevin O'Regan becomes particularly relevant. But also other E-approaches might prove their value in the study of phenomenal consciousness, as long as theorists don't slip back into thinking that their ultimate goal should be to find, from within the objective perspective, a causal explanation for the specific qualitative nature of our experience.

5.12 A misguided objection

Some believe that, in light of the HPC, assuming a strict physical-phenomenal identity relation fairs no better than other approaches. The reason is that it fails to account for why certain physical event-structures exhibit phenomenal consciousness, but other similar structures do not. Recently, Valerie Hardcastle has expressed such a worry with regard to Chemero's and Silberstein's proposed

identification of brain-body-environment interaction with cognition and phenomenal consciousness[115]:

> *Prima facie*, there is nothing about being an animal synergy that should give rise to conscious experience. In particular, there does not seem to [sic] anything special about an animal brain-body-environment interaction that an animal brain piece-body piece-environment interaction would not also share. (Hardcastle 2017: 6)

Her point is that assuming a strict physical-phenomenal identity does not escape the Hard Problem at all, and that an identity theory is still haunted by the same question the computationalist is faced with: "*Prima facie*, there is nothing about the computations themselves that should give rise to conscious experience, and there are certainly many computational systems that we believe are not conscious." (Hardcastle 2017: 6). We might add that this problem arises for the proponent of a supervenience or an emergence relation as well (or any other proposed dependency relation from the phenomenal to the physical). Why does phenomenal consciousness supervene on, or emerge from this physical substrate, rather than another? In light of the above, it should be clear that Hardcastle misses the point when she thinks the question of why certain brain-body-environment interactions can "give rise to conscious experience" when one assumes a strict identity between both. The whole idea of assuming identity is precisely to avoid questions of physical entities 'giving rise' to conscious experience. Hardcastle simply fails to appreciate the meaning of strict identity in this context. Her claim, therefore, that physical-phenomenal identity "does not solve the hard problem of consciousness" (Hardcastle 2017: 6) is simply beside the question. Assuming identity isn't supposed to solve the HPC, it is supposed to *dissolve* it. Nevertheless, I think Hardcastle's worry isn't entirely unjustified either. I'll return to this in the conclusion.

[115] See Chemero 2009. See also Silberstein and Chemero 2011 and 2015.

5.13 Concluding remarks

Assuming a strict identity relation between the phenomenal and the physical comes at a price. But if it manages to neutralize the Hard Problem of Consciousness, it is a price worth paying. Perhaps the theory's biggest cost is entailed by the fact that it requires us to accept – perhaps counterintuitively – the impossible nature of the Hard Problem. If the subjective qualitative nature of experience is truly identical to what from the objective perspective shows up as situated bodily (including neural) activity, then there no longer is any question of how to explain the former in terms of the latter (or vice versa, for that matter). For this to be possible, there would have to be an objective ontological difference between the two relata of the identity relation. But this is something that a strict identity relation does, of course, not allow. The two relata only exist as relative to our two different perspectives, not as an objective distinction in the world. Accepting identity, and thereby giving up on the idea that we can explain the phenomenal in terms of the physical, might be felt to be too counterintuitive to be plausible. As Smart once put it:

> That everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together-roughly, biology is to physics as radio-engineering is to electromagnetism) except the occurrence of sensations seems to me to be frankly unbelievable. (Smart 1959: 142)

In light of the above, perhaps the most apt response to Smart would be to correct his first sentence and instead say that not everything *should* be explicable in terms of physics. Identities, for instance, should not. However, it is important to stress that all of the above doesn't mean that phenomenal consciousness is something that principally eludes *all* explanation or that it is entirely miraculous. True, the assumed fact that some physical event-structures are identical with phenomenality means *ex hypothesi* that the physical does not cause the qualitative nature of experience. However, to the extent that the phenomenal is identical with something that can be studied as a causal phenomenon from the objective perspective, causal explanations in that field will also be relevant for our understanding of experiential phenomena. On the identity theoretical account, it

is impossible to explain in causal physical terms why it is that a toothache feels like a toothache, but on the assumption that the toothache experience is identical with certain situated bodily events, a better causal understanding of the latter should also help explain why someone has this specific sensation, and not some other experience (or no experience at all). Of course, this is easier said than done. For although it manages to escape the HPC's gravitational pull, for many, identity theory has a few recalcitrant problems of its own. From an empirical point of view, perhaps the most pertinent question is why it is that certain physical event-structures (dynamical situated bodily activity) are, as a matter of fact, experienced, while others are not. Of course, we can no longer expect to account for these identities (or non-identities) causally, for this makes no sense. From an identity theoretical point of view, the physical does not "give rise" to the phenomenal, as Hardcastle still puts it. Yet, we might nevertheless still legitimately wonder what it is about certain physical event-structures, but not others, that make these events identifiable, not *with*, but *as* an experience. To be clear, this is not the same as asking, nonsensically, for an explanation of the strict identity relation. Compare, for instance, the question "What causes a collection of H_2O molecules to be identical with water?" to the question "What is it about this liquid, but not another, that makes it identifiable as water?" On the assumption of a strict identity between water and H_2O, the first question is impossible to answer because it makes no sense. But the second question is not. It is answerable, and we know what the answer is: "The fact that the liquid is identical *with* a collection of H_2O molecules". In other words, from an empirical point of view, we might still want to know which physical event-structures can be identified as phenomenal consciousness, and how this can be determined. Indeed, assuming strict phenomenal-physical identity does not absolve us from these questions (which are questions for probably *all* mind-body theories). However, compared to the Hard Problem of Consciousness, these questions are probably in fact just *hard*, and not downright *impossible*.

References

Block, N. (2006). Philosophical issues about consciousness. In Nadel, L. (Ed.), *The Encyclopedia of Cognitive Science*: Wiley Online Library.

Broad, C.D.(1925). *The Mind and its Place in Nature*. New York: The Humanities Press Inc, London: Routledge & Kegan Paul LTD.

Chalmers, D. (1996). *The Conscious Mind*, New York: Oxford University Press.

Chemero, A. (2009). *Radical embodied cognition*. Cambridge: The MIT Press.

Crick, F. & Koch, C. (1990). Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences 2*: 263–275.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of The American Philosophical Association 60*: 441–458.

Davidson, D. (1995). Relations and transitions. An interview with Donald Davidson. *Dialectics 49*: 75–86.

Edelman, G. (1989). *The Remembered Present*. New York: Basic Books.

Feigl, H. (1958). The "Mental" and the "Physical"', in H. Feigl, M. Scriven and G. Maxwell (Eds.), *Concepts, Theories and the Mind-Body Problem* (Minnesota Studies in the Philosophy of Science, Volume 2). Minneapolis: University of Minnesota Press; reprinted with a Postscript in Feigl 1967.

Gallagher, S. & Zahavi, D. (2008). *The Phenomenological Mind*. London: Routledge.

Goff, P., Seager, W. and Allen-Hermanson, S. (2017). Panpsychism. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/panpsychism/>.

Goodman, N. (1978). *Ways of Worldmaking*. Indianapolis: Hackett Publishing Company.

Hameroff, S. (1994). Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies 1*: 91–118.

Hardcastle, V. (1996). The why of consciousness: a non-issue for materialists. *Journal of Consciousness Studies 3*: 7–13.

Hare, R.M. (1952). *The Language* of *Morals.* Oxford: Oxford University Press.

Heil, J. (2013). *Philosophy of Mind. A Contemporary Introduction*, 3rd edition. New York and London: Routledge.

Hurley, S. & Noë, A. (2003). Neural plasticity and consciousness. *Biology and Philosophy 18***:** 131–168.

Hutto, D.D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA.: MIT Press.

Huxley, T.H & Youmans, W.J. (1868). *The Elements of Physiology and Hygiene: A Text-book for Educational Institutions*. New York: D. Appleton and company.

Jackson, F. (1982). *Epiphenomenal Qualia*. *Philosophical Quarterly 32*: 127–136.

Kim, J. (1993) (Ed.) *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Kirchhoff, M. & Hutto, D.D. (2016). Never Mind the gap: neurophenomenology, radical enactivism and the hard problem of consciousness. *Constructivist Foundations 11*: 302–309.

Leibniz, G.W. (1714/1991). *Monadologie*. Trans. by N. Rescher. Pittsburgh, PA: University of Pittsburg Press.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly 64*: 354–436.

Llinas, R. & Ribary, U. (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences USA 90***:** 2078–2081.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*, Paris: Gallimard.

Merleau-Ponty, M. (1945/1948). La querelle de l'existentialisme. In *Sens et non-sens*. Paris: Editions Nagel.

Merleau-Ponty, M. (1964). *The battle over existentialism.* In *Sense and Non-sense*. (Hubert L. Dreyfus & Patricia Allen Dreyfus, Trans.) Evanston, Illinois: Northwestern University Press.

Myin, E. (2016). Perception as something we do. *Journal of Consciousness Studies 23*: 80–104.

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

Pepper, S.C .(1926). Emergence. *The Journal of Philosophy 23*: 241–245.

Place, U.T. (1956). Is consciousness a brain process? *British Journal of Psychology 47*: 44–50.

Prinz, J. (2006). Putting the brakes on enactive perception. *Psyche 12*: 1–19.

Rosenberg, A. (2011). *The Atheist's Guide to Reality: Enjoying Life without Illusions.* W.W. Norton & Company Inc.

Sellars, W. (1963). Philosophy and the scientific image of man. In Wilfrid Sellars, *Empiricism and the Philosophy of Mind*, London: Routledge & Kegan Paul Ltd: 1–40.

Silberstein, M.& Chemero, A. (2011). Complexity and extended phenomenological–cognitive systems. *Topics in Cognitive Science 4*: 35–50.

Silberstein, M. & Chemero, A. (2015). Extending neutral monism to the hard problem. *Journal of Consciousness Studies 22*: 181–194.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology 55***:** 349–374.

Singer, W. & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neurosciences 18***:** 555–586.

Smart, J.J.C. (1959). Sensations and brain processes. *Philosophical Review 68*: 141–156.

Smart, J.J.C. (1963). *Philosophy and Scientific Realism*. London: Routledge and Kegan Paul.

Smart, J. J. C. (2017). The mind/brain identity theory. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). URL = <https://plato.stanford.edu/archives/spr2017/entries/mind-identity/>.

Strawson, P.F. (1985). *Skepticism and Naturalism*. London: Methuen & Co. Ltd.

van Brakel, J. (1996). Interdiscourse or supervenience relations: the priority of the manifest image. *Synthese 106*: 253–297.

van Brakel, J. (1999). Supervenience and anomalous monism. *Dialectica 53*: 3–24.

Walter, S. (2002). Terry, Terry, quite contrary. *Grazer Philosophische Studien 63*: 103–122.

Epilogue

A few years ago, physicist Stephen Hawking incurred the wrath of the philosophical community by proclaiming in one of his books that "philosophy is dead." The opening lines read as follows:

> How can we understand the world in which we find ourselves? What is the nature of reality? How does the universe behave and why does it exist? Does it need a creator? Most of us don't worry about these questions most of the time. But almost all of us must sometimes wonder "Why are we here?", "Where do we come from?". Traditionally, these are questions for philosophy, but philosophy is dead. (Hawking & Mlodinow 2010: 3)

The recently deceased scientist leaves little doubt as to the cause of death, for he adds: "Philosophers have not kept up with modern developments in science, particularly physics." Otherwise put, philosophy has made itself irrelevant and, as Hawking further argues, it should entrust its inheritance to the much more capable hands of the physicist.

This portrayal of philosophy is a farcical mischaracterization. By suggesting that philosophy's life depends on its ability to offer answers to questions like the ones above, Hawking's criticism backfires, for he himself might be reproached for not having "kept up with modern developments" in philosophy.

First of all, the term 'philosophy' nowadays captures a rich, diversified field with countless subdomains, each with their own specializations. Philosophy is in this respect very similar to science. So saying that philosophy is dead because speculative metaphysics can't tell us 'why we are here' is like saying that science is dead because physics can't explain the Second World War.

Second, perhaps there was a time when metaphysical questions about divine creators or the meaning of life made up a substantial part of the philosophy student's curriculum, but today, the vast majority of academic philosophers does

not consider answering these questions their main occupation. In fact, they are by many philosophers held to be the wrong kind of questions to begin with. So, oddly enough, it is actually Hawking, not the philosopher, who still sees it as his intellectual duty to come up with definite answers to these perennial questions. For despite Hawking's claim that philosophy is dead, at the same time, these age-old philosophical questions are for the physicist still alive and kicking. It is, after all, the main ambition of his book to finally give a scientific answer to precisely these questions that have haunted metaphysics for ages.

In other words, it is not that Hawking criticizes philosophy for not having asked the right questions. The questions were right, the answers, however, were not. Fortunately, according to Hawking, we have good reason to assume that his Theory-M will finally provide us with the answer to all our questions, including philosophical classics, such as "Why is there something rather than nothing?" and "Why do we exist?" (Hawking & Mlodinow 2010: 10). Yet, how can philosophy be said to be dead when Hawking himself keeps resurrecting the same old philosophical questions?

The reader might begin to wonder why I'm mentioning all of this. Surely, I do not want to end my philosophical dissertation with a discussion of Hawking's ambivalent attitude towards philosophy. My reason for bringing it up is this: if there is one thing that my four year research has taught me, it is that, at least when it comes to the subjects of mind and cognition, philosophy is more necessary than ever. During my research, it has become increasingly clear to me that, with regard to the further development of cognitive science, philosophy should be – and actually *is* – playing a crucial role. This role is not so much that of providing potential answers to the cognitive scientist's questions. Rather, when it comes to understanding mind and cognition, philosophy's contribution lies first of all in questioning, not so much the answers or explanations of the cognitive scientist, but the questions that precede them. As Dennett correctly points out us:

> Every inquiry is in danger of setting off on the wrong foot, by asking the wrong questions. Wherever that happens, this is a job for philosophers! Philosophy—in every field of inquiry—is what you have to do until you figure out what questions you should have been asking in the first place. (Dennett 2013: 20)

This dissertation has tried to contribute to the difficult, but necessary task of 'asking the right questions'. Through a critical evaluation of some of the fundamental assumptions of today's prevalent theories of mind and cognition, I have tried to show that mainstream cognitive science has indeed set off on the wrong foot, and that this can only be remedied when it starts asking the right questions. Before learning how to run, cognitive science should first learn how to walk. Rather than devising yet another (but this time better) theory in terms of brains performing computational operations over content-carrying entities, perhaps questions such as "How does the brain compute?", "How is content stored in the brain?" or "How can we determine the format of internal representations?" should be traded in for less presumptuous questions, such as "Do brains compute?", "Does it make sense to think that semantic content can be stored in brains?", or "Do internal representations exist at all?". According to many working in the field, the answer to these questions is probably 'no'. The results of my research can be seen as supporting that answer.

This being said, however, I should emphasize that I do not want to suggest that cognitive scientists typically ask wrongheaded questions, whereas philosophers do not. Far from it. First, as we've seen in the second part of the dissertation, it is very likely that philosophical discussions about the mind-body relation and conscious experience should be understood as the result of asking the wrong questions. Rather than asking "How does the brain realize mental types?" or "How does conscious experience arise out of matter?", perhaps we should first wonder whether we are not in fact dealing with pseudo-problems resulting from wrong assumptions. Putting these questions into question, we should perhaps first ask "What is this realization relation?", or "Does conscious experience arise out of matter at all?" Second, as already indicated in the dissertation's introduction, in case of cognitive science theory, it is often hard to tell where the

philosophy ends and the science begins. Indeed, many of the core ideas of today's sciences of the mind find their origin in philosophy. Take mainstream cognitive science's cornerstone notion of mental representation, for instance. Although 'mental representation' is nowadays considered to be in the first instance a theoretical construct of cognitive science, the notion has a long philosophical tradition, going back at least to medieval philosophy. Or consider, as another example, the notion of mental content, a notion without which the idea of mental representation would become vacuous. Questions regarding the nature of content where already plaguing philosophers long before it got introduced into cognitive psychology. So if mainstream cognitive science might be said to have set off on the wrong foot by asking the wrong questions, this is partially the result of the wrong questions philosophers have been asking in the past. To be sure, philosophy's list of wrong questions is long. I can only hope that the questions I've raised in this dissertation do not add to its length.

General Bibliography

Adams, F. (2010a). Embodied cognition. *Phenomenology and the Cognitive Sciences: Special Issue on 4E Cognition 9*: 619–628.

Adams, F. (2010b). Why we still need a mark of the cognitive. *Cognitive Systems Research 11,* 324–331

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell-Wiley.

Adams, F., & Beighley, S. (2011). The mark of the mental. In J. Garvey (Ed.), *The continuum companion to the philosophy of mind*. London: Continuum International Publishing Group, 54–72.

Adams, F., & Garrison, R. (2013). The mark of the cognitive. *Minds & Machines, 23*, 339–352.

Bakhurst, D. (1991). *Consciousness and Revolution in Soviet Philosophy*. Cambridge: Cambridge University Press.

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs 28*: 801-831.*Gender and Science: New Issues*. The University of Chicago Press.

Bechtel, W. & Mundale, J. (1999). Multiple realizability revisited: linking cognitive and neural states. *Philosophy of Science 66*: 175–207.

Bechtel, W. (1998). Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cognitive Science 22*: 295–318.

Beer, R. (1995). A dynamical systems perspective on agent-environment interactions.
Artificial Intelligence, *72*: 173–215.

Beer, R. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J. A. Meyer, J. Pollack, & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior,* 421–429. Cambridge, MA: MIT Press.

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

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Dordrecht: Kluwer.

Block, N. & Fodor, J. (1972). What psychological states are not. *Philosophical Review 81*: 159–81.

Block, N. (2006). Philosophical issues about consciousness. In Nadel, L. (Ed.), *The Encyclopedia of Cognitive Science*: Wiley Online Library.

Block, N. (ed.) (1980). *Readings in Philosophy of Psychology*, vol. 1. Cambridge, Mass.: Harvard University Press.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues 13*: 31–45.

Boghossian, P. (2005). Is meaning normative? In C. Nimtz and A. Beckermann (Eds.),*Philosophy–science–scientific philosophy*, 205–218. Paderborn: Mentis.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology 43*: 1–22.

Braddon-Mitchell, D., & Jackson, F. (1996). *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Brandom, R.B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge MA: Harvard University Press.

Broad, C.D. (1925). *The Mind and its Place in Nature*. New York: The Humanities Press Inc, London: Routledge & Kegan Paul LTD.

Buekens, F. (2014). *De transparantie van waarheid*. Acco: Leuven.

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria 21*: 67–84.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology & Philosophy 27*: 49–71.

Cash, M. (2008). Thoughts and oughts. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action 11*: 93–119.

Cash, M. (2009). Normativity is the mother of intention: Wittgenstein, normative practices and neurological representations. *New Ideas in Psychology 27*: 133–147

Chakravartty, A. (2010). Informational versus functional theories of scientific representation. *Synthese 172*: 197–213.

Chalmers, D. (1996). *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of*

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*: 625-647.

Chemero, A. (2009). *Radical Embodied Cognition*. Cambridge: The MIT Press.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. New York: Cambridge University Press.

Clark, A. (1996). *Being There*: *Putting Brain, Body and World together again*. MIT Press: Cambridge, MA.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences, 3*: 345–351.

Clark, A. (2005). Beyond the flesh: Some lessons from a mole cricket. *Artificial Life 11*: 233–244.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*: 401–431.

Craik, K.J.W. (1943/1967). *The Nature of Explanation*. Cambridge University Press.

Crick, F. & Koch, C. (1990). Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2: 263–275.

Currie, G. (1995). Visual imagery as the simulation of vision. *Mind and Language 10*: 25–44.

Currie, G. & Ravenscroft, I. (2003). *Recreative Minds*. Oxford: Oxford University Press.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of The American Philosophical Association*, 60: 441-458.

Davidson, D. (1995). Relations and transitions. An interview with Donald Davidson. *Dialectics,* 49: 75-86.

Davidson, D. (2005). *Truth and Predication*. Cambridge, Mass.: Belknap Press.

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese, 191*: 3639–3648.

Dennett, D. C. (1969). *Content and Consciousness*. Routledge and Kegan Paul plc.

Dennett, D.C. (1978). *Brainstorms*: *Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Dennett, D.C. (2013). *Intuition Pumps And Other Tools for Thinking.* W.W. Norton Compagny

Di Paolo, E. A., Buhrmann, T., & Barandiaran, X.E. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford University Press.

Dreyfus, H. L. (Ed.) (1982). *Husserl, Intentionality, and Cognitive Science.* MIT Press/Bradford Books, Cambridge.

Edelman, G. (1989). *The Remembered Present*. New York: Basic Books.

Edelman, S. (2003). But will it scale up? Not without representations. *Adaptive Behavior, 11*: 273–275.

Egan, F. (2012). Representationalism. In Margolis, E., Samuels, R. & Stich, S. (Eds.), *The Oxford Handbook of Philosophy and Cognitive Science*. Oxford University Press.

Epstein, R. (2016). The empty mind. Published online by Aeon. URL= https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Ervin-Tripp, S. (1967). An Issei learns English. *Journal of Social Issues 2*: 78–90.

Evans-Pritchard, E.E. *Nuer Religion*. Clarendon Press: Oxford. 1956

Feigl, H. (1958). The "Mental" and the "Physical"', in H. Feigl, M. Scriven and G. Maxwell (Eds.), *Concepts, Theories and the Mind-Body Problem* (Minnesota Studies in the Philosophy of Science, Volume 2). Minneapolis: University of Minnesota Press; reprinted with a Postscript in Feigl 1967.

Flament-Fultot, M. (2014). On genic representations. *Biological Theory 9*:149–162.

Fodor, J. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive science. *The Behavioral and Brain Sciences 3*: 63–73.

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

Fodor, J. (1987). Meaning and the world order. In *Psychosemantics*, 97–133. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, Z. (1981). How direct is visual perception? Some reflections on Gibson's 'ecological approach'. *Cognition 9*: 139-196.

Foucault, M. (1972/1980). Truth and power: an interview with Michel Foucault. In C. Gordon (Ed.), *Power/Knowledge: Selected Interviews & Other Writings 1972-1977*. New York: Pantheon Books.

French, S. (2003). A model-theoretic account of representation (or, I don't know much about art…but I know it involves isomorphism). *Philosophy of Science 70*: 1472–1483.

Gallagher, S., & Zahavi, D. (2008). *The Phenomenological Mind*. London: Routledge.

Gallagher, S., & Miyahara, K. (2012). Neo-pragmatism and enactive intentionality. In: Schulkin, J. (Ed.) *Action, Perception and the Brain. New Directions in Philosophy and Cognitive Science*. Palgrave Macmillan: London

Gelbard-Sagiv, H. et al. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, *322*(5898), 96–101.

Gibbard, A. (2003). Thoughts and norms. *Philosophical Issues 13*: 83–98.

Gibbard, A. (2005). Truth and correct belief. *Philosophical Issues 15*: 338–350.

Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis 64*: 316–323.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric 40*: 63–90.

Godfrey-Smith, P. (2014). Signs and symbolic behavior. *Biological Theory 9*: 78-88.

Goff, P., Seager, W. & Allen-Hermanson, S. (2017). Panpsychism. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/panpsychism/>.

Goodman, N. (1968/1976): *Languages of Art: An Approach to a Theory of Symbols*. 2nd ed., Indianapolis: Hackett Publishing Company.

Goodman, N. (1978). *Ways of Worldmaking*. Indianapolis: Hackett Publishing Company.

Hameroff, S. (1994). Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies 1*: 91–118.

Hardcastle, V. (1996). The why of consciousness: a non-issue for materialists. *Journal of Consciousness Studies 3*: 7-13.

Hardcastle, V. (2017). https://doi.org/10.1007/s11245-017-9503-7

Hare, R.M. (1952). *The Language* of *Morals.* Oxford: Oxford University Press.

Hattiangadi, A. (2006). Is meaning normative? *Mind & Language 21*: 220–240.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory*. Hillsdale, N.J.: Erlbaum.

Haugeland, J. (1998). *Having thought: Essays in the metaphysics of mind*, 305–361. Cambridge, MA: Harvard University Press.

Hawking, S., & Mlodinow, L. (2010). *The Grand Design*. New York: Bantam Books.

Heil, J. (1992). *The Nature of True Minds*. New York: Cambridge University Press.

Heil, J. (2013). *Philosophy of Mind. A Contemporary Introduction*, 3rd ed. New York and London: Routledge.

Hume, D. (2002/1739-40). *A Treatise of Human Nature*. 2nd ed. Norton, D.F., & Norton, M.J. (Eds.) Oxford: Oxford University Press.

Hurley, S. (1998). Vehicles, contents, conceptual structure, and externalism. *Analysis 58*: 1-6.

Hurley, S. & Noë, A. (2003). Neural plasticity and consciousness. *Biology and Philosophy 18*: 131–168.

Husserl, E. (2003). *Transzendentaler Idealismus. Texte aus dem Nachlass (1908–1921)*. Husserliana 36. R. Rollinger (Ed.). Dordrecht: Kluwer Academic Publishers.

Hutto, D. D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Hutto, D. D. (2009). Mental representation and consciousness. In W. P. Banks (Ed.), *Encyclopedia of Consciousness 2*: 19-32. Oxford: Elsevier.

Hutto, D.D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA.: MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

Hutto, D.D. ,& Myin, E. (forthcoming). Deflating deflationism about mental representation. In T. Schlicht et al. (Eds.), *Mental Representation*. Oxford: Oxford University Press.

Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia 43*(3), 521–536.

Hutto, D.D., Myin, E., Peeters, A., & Zahnoun, F. (forthcoming). The cognitive basis of computation: putting computation in its place. In Sprevak, M. & Colombo, M. (Eds.), *The Routledge Handbook of the Computational Mind*. Routledge.

Huxley, T.H & Youmans, W.J. (1868). *The Elements of Physiology and Hygiene: A Text-book for Educational Institutions*. New York: D. Appleton and company.

Jackson, F. (1982). *Epiphenomenal Qualia*. *Philosophical Quarterly 32*: 127–136.

James, W. (1907). Pragmatism and humanism. Lecture 7. In *Pragmatism: A new name for some old ways of thinking*. New York: Longman Green & Co.

Kalish, C. (2005). Becoming status conscious: Children's appreciation of social reality. *Philosophical Explorations 8*: 245–263.

Kim, J. (1972). Phenomenal properties, psychophysical laws and identity theory. *Monist* 56: 177–192.

Kim, J. (1993) (Ed.) *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Kirchhoff, M., & Hutto, D.D. (2016). Never Mind the gap: neurophenomenology, radical enactivism and the hard problem of consciousness. *Constructivist Foundations 11*: 302–309.

Kirsh, D. (2010). Thinking with external representations. AI & Society *25*: 441–454.

Kosslyn, St. M. (1978). Imagery and Internal Representation. In Rosch, E. & Lloyd, B.B. (Eds.), *Cognition and Categorization.* Hillsdale: Lawrence Erlbaum Associates, 217-257.

Kosslyn, St. M. (1980). *Image and Mind.* Cambridge, Mass.: Harvard University Press.

Kosslyn, St. M., Thompson, W. L., & Ganis, G. (2010). *The Case for Mental Imagery.* Oxford University Press.

Leibniz, G.W. (1714/1991). *Monadologie*. Trans. by N. Rescher. Pittsburgh, PA: University of Pittsburg Press.

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly 64*: 354–436.

Lewis, D. K. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Llinas, R. & Ribary, U. (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences USA 90***:** 2078–2081.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior 13*: 585–589.

Luntley, M. (1999). *Contemporary Philosophy of Thought: Truth, World, Content.* Oxford: Blackwell Publishers Ltd.

Lycan, W., (1974). Kripke and the Materialists. *Journal of Philosophy 71*: 667-689.

Mackenzie, J.L. (2004). Semantic categories and operations in morphology I: entity concepts. In: Geert Booij et al. (Eds.), *Morphology. An International Handbook on Inflection and Word-Formation*. Vol 2. Berlin: de Gruyter, 973–983.

Manzotti, R., & Pepperell, R. (2013). Denying the content–vehicle distinction: a response to'The New Mind Revisited'. *AI and Society 28*: 467–470.

Marbach, E. (1993). *Mental Representation and Consciousness: Towards a Phenomenological Theory of Representation and Reference*. Dordrecht: Kluwer Academic Publishers.

McDowell, J.H. (1998). *Mind, Value and Reality*. Cambridge, MA: Harvard University Press.

McGinn, C. (1989). *Mental Content.* Oxford: Basil Blackwell.

McIntyre, R. (1986). Husserl and the Representational Theory of Mind. *Topoi 5*: 101–113**.**

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*, Paris: Gallimard.

Merleau-Ponty, M. (1964). *The battle over existentialism.* In *Sense and Non-sense*. (Hubert L. Dreyfus & Patricia Allen Dreyfus, Trans.) Evanston, Illinois: Northwestern University Press.

Merleau-Ponty, M. (1945/1948). La querelle de l'existentialisme. In *Sens et non-sens*. Paris: Editions Nagel.

Miłkowski, M. (2015). The hard problem of content: Solved (long ago). *Studies in Logic, Grammar and Rhetoric 41*: 73–88.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA.: MIT Press.

Millikan, R. G. (1993a). *White Queen Psychology and Other Essays for Alice*. Cambridge, MA.: MIT Press.

Millikan, R. G. (1993b). Content and vehicle. In N. Eilan, R. McCarthy, B. Brewer (Eds.). *Spatial Representation*. Oxford: Blackwell. 256–268.

Morgan, A., & Piccinini, G. (2017). Towards a Cognitive Neuroscience of Intentionality. *Minds and Machines*. Doi. 10.1007/s11023-017-9437-2.

Myin, E. (2016). Perception as something we do. *Journal of Consciousness Studies 23*: 80–104.

Myin, E. & Loughlin, V. (in press). Sensorimotor and enactive approaches to consciousness. In Gennaro, R. (Ed.), *Routledge Handbook of Consciousness*.

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery 19*: 113–126.

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

O'Brien, G., & J. Opie (2015). Intentionality lite or analog content? *Philosophia 43*: 723–730.

Orlandi, N. (2014). *The Innocent Eye*: *Why Vision is not a Cognitive Process*. New York: Oxford University Press.

Papineau, D. (1998) Mind the gap. *Philosophical Perspectives*, *12*: 373–89

Pepper, S.C .(1926). Emergence. *The Journal of Philosophy 23*: 241–245.

Pessoa, L., Thompson, E., and Noë, A. (1998). Finding out about filling-in: a guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences 21*: 723-802.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University

Pitt, D. (2017). Mental Representation. *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/spr2017/entries/mental-representation/.

Place, U.T. (1956). Is consciousness a brain process? *British Journal of Psychology 47*: 44–50.

Poland, J. (1994). *Physicalism: The Philosophical Foundations*. New York: Oxford University Press.

Polger, T. (2007). Realization and the Metaphysics of Mind. *Australasian Journal of Philosophy 85*: 233–59.
Polger, T. (2002). Putnam's intuition. *Philosophical Studies 109*: 143–170.
Polger, T. (2004). *Natural Minds*. Cambridge, MA: MIT Press.
Polger, T. (2009). Evaluating the Evidence for Multiple Realization. *Synthese 167*: 457–472.
Polger, T. (2013). Realization and multiple realization, chicken and egg. *European Journal of Philosophy 23*: 862–877.
Polger, T., & Shapiro, L. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press.
Preston, B. (1994). Husserl's non-representational theory of mind. *The Southern Journal of Philosophy 32*: 209–232.
Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
Prinz, J. (2006). Putting the brakes on enactive perception. *Psyche 12*: 1–19.
Puccetti, R. (1977). The great C-fiber myth: a critical note. *Philosophy of Science 44*: 303–305.
Putnam, H. (1975). *Mind, Language, and Reality: Philosophical Papers*, Volume 2. Cambridge: Cambridge University Press.
Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.
Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin, 80*: 1–24.
Pylyshyn, Z. (1980). Computation and cognition: issues in the foundation of cognitive science. *The Behavioral and Brain Sciences 3*: 111–132.
Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.
Ramsey, W. M. (2007). *Representation Reconsidered.* New York: Cambridge University Press.
Ramsey, W. M. (2017). Must cognition be representational? *Synthese, 194*: 4197-4214.
Ramsey, W.M. (forthcoming). Defending representation realism. In T. Schlicht et al. (Eds.), *Mental Representations.* Oxford University Press.
Rescorla, M. (2016). Bayesian Sensorimotor Psychology. *Mind & Language 31*: 3–36.
Reynaert, P. (2015) Does naturalism commit a category mistake? *Bulletin d'Analyse Phénoménologique 9*: 1–20.
Rosenberg, A. (2011). *The Atheïst's Guide to Reality: Enjoying Life without Illusions.* W.W. Norton & Company, Inc.: New York, London.
Rowlands, M. (2006). *Body Language: Representation in Action*. MIT Press.
Rowlands, M. (2010). *The New Science of the Mind*. Cambridge, MA: MIT Press.

Rowlands, M. (2014). Arguing about representation. *Synthese.* Doi:10.1007/s11229-014-0646-4

Rozin, P., & Nemeroff, C. (2002). Sympathetic magical thinking: the contagion and similarity "heuristics". In T. Gilovich, D. Griffin & D. Kahnemann (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Ryle, G. (1949/2009). *The Concept of Mind*. London: Hutchinson.

Sellars, W. (1962). Truth and "correspondence". *The Journal of Philosophy*, *59*: 29–56.

Sellars, W. (1963). Philosophy and the scientific image of man. In Wilfrid Sellars, *Empiricism and the Philosophy of Mind*, London: Routledge & Kegan Paul Ltd: 1–40.

Shapiro, L. (2000). Multiple realizations. *Journal of Philosophy 97*: 635–654.

Shapiro, L. (2004). *The Mind Incarnate*. Cambridge, MA: MIT Press.

Shapiro, L. (2008). How to test for multiple realization. *Philosophy of Science 75*: 514–525.

Shapiro, L. (2011). *Embodied Cognition.* London: Routledge.

Shapiro, L. (2014). *Radicalizing Enactivism: Basic Minds without Content*, by Daniel D. Hutto & Erik Myin (Review). *Mind 123*: 213–220.

Shea, N. (2013). Naturalising representational content. *Philosophical Compass 8*: 496–509.

Silberstein, M., & Chemero, A. (2011). Complexity and extended phenomenological–cognitive systems. *Topics in Cognitive Science 4*: 35–50.

Silberstein, M., & Chemero, A. (2015). Extending neutral monism to the hard problem. *Journal of Consciousness Studies 22*: 181–194.

Singer, W. & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neurosciences 18***:** 555–586.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology 55***:** 349–374.

Skyrms, B. (2010). *Signals: Evolution, Learning and Information.* Oxford: Oxford University Press.

Smart, J.J.C. (1959). Sensations and brain processes. *Philosophical Review 68*: 141–156.

Smart, J.J.C. (1963). *Philosophy and Scientific Realism*. London: Routledge and Kegan Paul.

Smart, J.J.C. (2017). The mind/brain identity theory. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). URL = <https://plato.stanford.edu/archives/spr2017/entries/mind-identity/>.

Strawson, P.F. (1985). *Skepticism and Naturalism: Some Varieties*. London: Methuen & Co. Ltd.

Strawson, P. F. (2008). *Freedom and Resentment and Other Essays.* London: Routledge.

Suarez, M. (2003). Scientific representation: against similarity and isomorphism. *International Studies in the Philosophy of Science 17*: 225-244.

Thomasson, A. L. (2013). Categories. In *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL <http://plato.stanford.edu/archives/fall2013/entries/categories/>.

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

Tonneau, F. J. (2011/2012). Metaphor and truth : a review of representation reconsidered by W. M. Ramsey. *Behavior and Philosophy (Online) 39/40:* 331-343.

van Brakel, J. (1996). Interdiscourse or supervenience relations: the priority of the manifest image. *Synthese 106*: 253–297.

van Brakel, J. (1999). Supervenience and anomalous monism. *Dialectica 53*: 3–24.

van Dijk, L. (2016). Laying down a path in talking. *Philosophical Psychology 29*: 993–1003.

van Dijk, L., & Withagen, R. (2016). Temporalizing agency: Moving beyond on- and offline cognition. *Theory & Psychology, 26*: 5–26.

van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy 92*: 345–381.

van Rooij, I., Bongers, R., & Haselager, W. (2002). A non-representational approach to imagined action. *Cognitive Science 26*: 345–375.

Vitrano, D. M. (2012). Comparing perception and imagination at the visual cortex. Dickinson College Honors Theses.

Walter, S. (2002). Terry, Terry, quite contrary. *Grazer Philosophische Studien 63*: 103–122.

Waskan, J. (2006). *Models and Cognition*. Cambridge MA: MIT Press.

Wheeler, M. (2005). *Reconstructing the Cognitive World.* Cambridge, MA: MIT Press.

Wiggins, D. (2001) *Sameness and Substance Renewed*. Cambridge: Cambridge University

Wilson, R. (2001). Two views of realization. *Philosophical Studies 104*: 1–30.

Wilson, R. A., & Foglia, L. (2011). Embodied cognition. *The Stanford Encyclopedia of Philosophy*. Zalta, E. N. (ed.), URL = http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/.

Wittgenstein, L. (2009/1953). *Philosophische Untersuchungen/ Philosophical Investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker & Joachim Schulte, Revised fourth edition by P. M. S. Hacker & Joachim Schulte. Wiley-Blackwell: West-Sussex.

Zahavi D. (2018). Brain, Mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism. *Husserl Studies, 34*: 47-61.

Zarkadakis, G. (2016). *In our own image.* UK: Rider Books.

Acknowledgements

It is said that writing a PhD dissertation is one of the most demanding and straining activities a person could engage in. Yet, although it has indeed sometimes been a struggle, looking back at these past four years, I am mainly left with a feeling of gratitude.

First of all, I am grateful to my supervisor, Erik Myin, for having trusted me with this wonderful research opportunity. But I also feel privileged for having had the chance of working with a supervisor whose philosophical ideas are in such close agreement with my own interests and intuitions. In the course of these four years, field experience has taught me that finding a philosophically kindred spirit isn't at all obvious.

I am also deeply indebted to my colleague, Karim Zahidi. His incomparably sharp mind has more than once detected flaws in my arguments, allowing me to improve them. I am sure that I have learned more from our countless discussions than I'm currently aware of.

Next, I want to thank my other colleagues Ludger van Dijk, Raoul Gervais and Jan Potters for their insights and their help with some of the chapters. In particular, however, I want to thank Jan van Eemeren, not only for being a very helpful colleague, but a great friend as well.

Much gratitude also goes to Daniel Hutto, my philosophical mentor in Australia. I want to thank him for granting me the opportunity of working at UOW's School of Humanities and Social Inquiry. Being in his company has taught me the true meaning of 'philosophical labor'. Here, I also want to thank my other Australian colleagues for making me feel welcome 'down under', in particular Michael Kirchhoff, Glenda Satne, Miguel Segundo Ortin and Anco Peeters.

Of course, this dissertation would not have been possible without the support of my friends, of whom I will only name a few here. I want to thank Koen Sels,

Veronik Willems, Jelle Dehaes, Rinus Van de Velde, Joyce de Badts, Maarten Gehem, Elli Bleeker, Anna Vanaerschot, Natasha Woolmer, Wim de Busser, Seppe Koop and Jan Zienkowski for having always been there throughout these past four years. I am, however, especially grateful for the warm kindness and support of Alexander Daems and Suzanne Grotenhuis, who have by now become like family to me. I am forever in their debt for providing me with, not just a room, but a home, and for helping me through some of the roughest moments of my life.

Yet, I am still more grateful for having met my wonderful girlfriend, Jozefien Van der Aelst. I can no longer imagine my life without her, but if it wasn't for this dissertation, our paths would never have crossed at the university hallway. I am sure, however, they would have crossed elsewhere.

Most of all, still, I want to thank the following three people: Arnold Burms, for making me a philosopher, and my grandfather and mother, Gustaaf and Els van de Gehuchte, for everything else. This dissertation is dedicated to her.

Abstract

The dissertation, titled *Mind, Mechanism and Meaning*, critically investigates two central assumptions of mainstream cognitive science and philosophy of mind: the commitment to the notion of internal representation on the one hand, and to the idea of the multiple realizability of the mental on the other. With regard to the notion of internal representation, the dissertation argues that this notion is ultimately untenable in that, to the effect that internal representations are understood as content-carrying vehicles with causal explanatory power, the notion is grounded in a confusion between the descriptive and the prescriptive/normative. The thesis is defended that *all* content-carrying entities, including representations, are socio-normatively constituted and should therefore be excluded from non-normative causal explanations of cognition. The results of the research support a non-representational approach to mind and cognition, as exemplified in various forms of E-Cognition, particularly in radical enactive/embodied approaches. Understanding human cognition requires taking into account the whole subject, that is, the subject as 'embrained', embodied, and embedded within an enacted normative intersubjective niche.

With regard to the idea of the multiple realizability of the mental, the dissertation argues that the idea can only be made intelligible against a particular metaphysical background, one that does not sit well with the intersubjective normative notions the idea of multiple realization conceptually relies on (types). Furthermore, it is argued that, even if we were to accept such a metaphysics, multiple realization is still not capable of providing the argument against identity theory which has come to be so widely accepted. The thesis is defended that there really is no strong argument against an identity theory, and that, in addition, assuming a strict identity between the mental and the physical is still a viable, perhaps even the *only* viable approach to the Hard Problem of Consciousness.

Samenvatting

Deze doctoraatsverhandeling, getiteld 'Cognitie, mechanisme en betekenis', wil een kritisch onderzoek zijn naar een aantal van de belangrijkste basisassumpties binnen de hedendaagse cognitieve wetenschappen enerzijds, en de filosofische psychologie anderzijds. Overeenkomstig dit onderscheid valt de dissertatie uiteen in twee delen.

Het eerste deel behandelt een centrale problematiek van de toonaangevende stroming binnen de cognitiewetenschappen, met name het probleem van interne representatie. De heersende opvatting hier is dat cognitie in de eerste plaats moet begrepen worden in termen van de manipulatie van interne representaties. In dit dominante representationalistische paradigma wordt het brein opgevat als een soort informatieverwerker, in veel opzichten gelijkend op, of zelfs letterlijk gelijk aan een digitale computer. Het brein heeft in dit model gewoonlijk de dubbele representationele functie van de externe werkelijkheid aan het organisme te 'beschrijven', en bepaalde gepaste acties aan het lichaam 'voor te schrijven'. Het zou dit doen via de constructie en manipulatie van – vermoedelijk neuraal gerealiseerde – interne representaties. Hoe deze entiteiten precies moeten begrepen worden is meestal niet helemaal duidelijk, maar in het algemeen worden zij gedefinieerd als materieel gerealiseerde structuren die op een of andere manier 'zeggen' hoe de wereld is. Interne representaties worden dus verondersteld een welbepaalde semantisch evalueerbare inhoud te hebben. Deze semantische evalueerbaarheid wordt dan verder begrepen in termen van het hebben van waarheidscondities of, meer algemeen, 'gepastheidscondities'. Dit betekent, met andere woorden, dat de notie van interne representatie een normatief karakter heeft.

Het eerste deel van deze dissertatie is hoofdzakelijk gericht op een evaluatie van deze centrale notie van interne representatie. Ik zal argumenteren dat deze verklarende basisnotie uiteindelijk onhoudbaar is en wat dit betekent voor bestaande verklaringen van cognitie binnen het huidige representationalistische

verklaringsmodel. Mijn argumentatie zal drie hoofdstukken in beslag nemen, hoofdstukken die in feite ook als zelfstandige artikels kunnen beschouwd worden.

Het eerste hoofdstuk, getiteld *Identifying Representations*, probeert eerst en vooral aan te tonen hoe, binnen bestaande theorieën van cognitie, vaak niet genoeg rekening gehouden wordt met het cruciale onderscheid tussen wat het betekent een representatie te 'zijn' enerzijds, en als een representatie 'gezien te kunnen worden' anderzijds. Onvoldoende rekening houden met dit essentiële onderscheid leidt, zoals zal aangetoond worden aan de hand van een concreet voorbeeld, tot een over-applicatie van representatie. Met andere woorden, hoewel cognitieve systemen vaak in termen van de manipulatie van representaties kunnen begrepen worden, betekent dit niet dat zij ook zo *moeten* begrepen worden, of dat het zelfs maar een goede manier zou zijn om deze systemen te begrijpen. Door dit onderscheid te miskennen worden potentiële niet-representationele verklaringen van cognitie bij voorbaat, en onterecht, uitgesloten. De tweede doelstelling van dit eerste hoofdstuk probeert, in aanloop naar het tweede hoofdstuk, duidelijk te maken hoe het specifiek normatieve karakter van representatie de technische notie van interne representatie voor wellicht onoplosbare problemen stelt.

Het tweede hoofdstuk, getiteld *Dereifying Representation*, bouwt op dit laatste verder door het welbepaalde normatieve karakter van representatie te onderzoeken. Zoals gezegd worden interne representaties gedefinieerd als materieel gerealiseerde structuren met een welbepaalde semantische inhoud. Omdat representaties hier begrepen worden als materiële objecten zal ik hier spreken van de gereïficeerde opvatting van representatie. Eerst zal aangetoond worden dat de gereïficeerde notie van representatie onhoudbaar is vanwege het specifiek sociaal-normatieve karakter van de soort semantische inhoud die interne representaties geacht worden te bezitten. Representatie is geen object, het is een sociaal-normatief geconstitueerd fenomeen dat simpelweg niet bestaat buiten een sociaal-normatieve context. Vervolgens zal duidelijk gemaakt worden

welke factoren een rol spelen bij zowel de totstandkoming, als de instandhouding van deze ongefundeerde tendens tot reïficatie.

Hoofdstuk drie, getiteld *Off-line Cognition as Representation Hungry?* gaat dieper in op één van de meest centrale discussies binnen de huidige filosofie van de cognitieve wetenschap. Een vaak gehoorde kritiek aan het adres van theoretici die cognitie proberen te benaderen vanuit een niet-representationalistische hoek is dat deze benadering noodzakelijk explanatorisch gelimiteerd is. Het door de meerderheid aanvaarde idee lijkt nu te zijn dat, wanneer het gaat over het verklaren van bepaalde 'on-line' vormen van cognitie, interne representaties misschien overbodig zijn. Verklaren hoe een organisme in staat is om zichzelf in evenwicht te houden, bijvoorbeeld, kan waarschijnlijk beter zonder representaties gebeuren. Wanneer het echter aankomt op de verklaring van meer 'echt' mentale fenomenen zoals herinneren, verbeelden, denken en redeneren, blijft volgens de meeste theoretici het inroepen van interne representaties een *conditio sine qua non*. Ik zal trachten te beargumenteren dat deze principiële afbakening berust op een verwarring tussen het niveau van de beschrijving en het niveau van de verklaring en dat we bij nader inzien geen goede redenen hebben om aan te nemen dat hogere vormen van cognitie ('off-line' cognitie) moeten verklaard worden in termen van de constructie en manipulatie van interne, semantische-inhoudsdragers (representaties).

Zoals aangegeven zal in het tweede deel overgeschakeld worden naar een thematiek die eerder thuishoort binnen de filosofische psychologie (Philosophy of Mind). Meer bepaald zal gekeken worden naar het invloedrijke idee van 'meervoudige realiseerbaarheid van het mentale', en hoe het zich verhoudt tot een potentiële lichaam-geest identiteitstheorie. De bespreking moet begrepen worden tegen de achtergrond van het klassiek filosofische lichaam-geest probleem. Dit tweede deel van de dissertatie bestaat zelf uit twee hoofdstukken (vier en vijf) die eveneens als onafhankelijke artikels kunnen gelezen worden.

Hoofdstuk vier, getiteld *Multiple Realization: A Thesis with Identity Issues*, behandelt kritisch de wijdverspreide functionalistische opvatting dat het mentale

meervoudig gerealiseerd kan worden in het fysische. Deze opvatting wordt bovendien aanvaard als een doorslaggevend argument tegen een potentiële identiteitstheorie. Als één en hetzelfde mentale type gerealiseerd kan worden door verschillende types van fysische structuren, dan kan het mentale onmogelijk strikt identiek zijn met het fysische. Ik zal trachten aan te tonen dat het argument van meervoudige realiseerbaarheid tegen identiteitstheorie berust op een onterechte samenvoeging van twee zeer verschillende identiteitsbegrippen. Tevens zal het empirische statuut van de these van meervoudige realiseerbaarheid in vraag gesteld worden, te meer omdat zij berust op een aantal begrippen ('types', 'realizability') waarvan onduidelijk is hoe zij naturalistisch moeten begrepen worden.

Het vijfde en laatste hoofdstuk tenslotte (*Reconsidering Identity*) gaat, na de identiteitstheorie vrijgepleit te hebben van de onterechte beschuldigingen van het 'meervoudige realiseerbaarheidsargument', dieper in op de mogelijkheid van een strikte lichaam-geest identiteit. Er zal aangetoond worden dat er uiteindelijk geen goede argumenten zijn om een correct begrepen (i.e., niet reductionistische) identiteitstheorie uit te sluiten. Dit is belangrijk in het licht van het zogenaamde 'Harde Probleem van Bewustzijn' (Hard Problem of Consciousness). Zolang het subjectieve fenomenaal bewustzijn begrepen wordt in een causale afhankelijkheidsrelatie met een fysisch substraat blijft het 'harde probleem' van hoe subjectieve ervaring kan ontstaan uit louter fysische processen zich stellen. Een veronderstelde identiteitsrelatie heeft dit probleem niet. Wanneer we mogen aannemen dat, wat vanuit één perspectief als fenomenaal bewustzijn kan gekarakteriseerd worden, strikt identiek is met wat vanuit een ander perspectief verschijnt als fysische spatiotemporele structuren, dan wordt het Harde Probleem van Bewustzijn geneutraliseerd. Vanuit de optiek van een identiteitstheorie verwordt het probleem dus tot een pseudoprobleem. Immers, tussen twee schijnbaar verschillende, maar in realiteit identieke zaken, kan geen causale relatie bestaan. Deze moet dus ook niet verklaard worden.

Colophon