Peter Fazekas

Tagging the World: Descrying Consciousness in Cognitive Processes

PhD Thesis

School of Philosophy, Psychology and Language Sciences
The University of Edinburgh
2012

Word Count

Full:	**116 524**
Preliminaries:	3 435
Main text:	91 564
Footnotes:	16 260
Bibliography:	5 265

Table of Contents

PART 1: SETTING THE STAGE

Chapter 1: Physicalism

PART 2: THE MONADIC MARKER ACCOUNT

Chapter 4: Monadic Markers

Declaration

Hereby, I declare that this PhD thesis is my own work, and has not been submitted for any other professional degree or qualification.

Peter Fazekas

Note on Publication

Parts of Chapter 5, Chapter 6, and Chapter 7 have been published, with support from Prof. Jesper Kallestrup.

Parts of Chapter 5 (§5.2) were published as:

Cognitive Architecture and the Epistemic Gap: Defending Physicalism without Phenomenal Concepts (2011). *Philosophia*, *39*(1), 21-29.
DOI: 10.1007/s11406-010-9274-5

Parts of Chapter 6 (§6.2) and parts of Chapter 7 (§7.2.2) were published as:

Reconsidering the Role of Bridge Laws in Inter-Theoretical Reductions (2009). *Erkenntnis*, *71*(3), 303-322.
DOI: 10.1007/s10670-009-9181-y

Parts of Chapter 7 (§7.2.1) were published as:

Causation at Different Levels—Tracking the Commitments of Mechanistic Explanations (2011). [with Gergely Kertész] *Biology and Philosophy*, *26*(3), 365-383.
DOI: 10.1007/s10539-011-9247-5

Abstract

Although having conscious experiences is a fundamental feature of our everyday life, our understanding of what consciousness is is very limited. According to one of the main conclusions of contemporary philosophy of mind, the qualitative aspect of consciousness seems to resist functionalisation, i.e. it cannot be adequately defined solely in terms of functional or causal roles, which leads to an epistemic gap between phenomenal and scientific knowledge. Phenomenal qualities, then, seem to be, in principle, unexplainable in scientific terms. As a reaction to this pessimistic conclusion it is a major trend in contemporary science of consciousness to turn away from subjective experiences and re-define the subject of investigations in neurological and behavioural terms. This move, however, creates a gap between scientific theories of consciousness, and the original phenomenon, which we are so intimately connected with.

The thesis focuses on this gap. It is argued that it is possible to explain features of consciousness in scientific terms. The thesis argues for this claim from two directions. On the one hand, a specific identity theory is formulated connecting phenomenal qualities to certain intermediate level perceptual representations which are unstructured for central processes of the embedding cognitive system. This identity theory is hypothesised on the basis of certain similarities recognised between the phenomenal and the cognitive-representational domains, and then utilised in order to uncover further similarities between these two domains. The identity theory and the further similarities uncovered are then deployed in formulating explanations of the philosophically most important characteristics of the phenomenal domain—i.e. why phenomenal qualities resist functionalisation, and why the epistemic gap occurs.

On the other hand, the thesis investigates and criticises existing models of reductive explanation. On the basis of a detailed analysis of how successful scientific explanations proceed a novel account of reductive explanation is proposed, which utilises so-called prior identities. Prior identities are prerequisites rather than outcomes of reductive explanations. They themselves are unexplained but are nevertheless necessary for mapping the features to be explained onto the features the explanation relies on. Prior identities are hypothesised in order to foster the formulation of explanatory claims accounting for target level phenomena in terms of base level processes—and they are justified if they help projecting base level explanations to new territories of the target level.

The thesis concludes that the identity theory proposed is a prior identity, and the explanations of features of the phenomenal domain formulated with the aid of this identity are reductive explanations proper. In this sense, the thesis introduces the problem of phenomenal consciousness into scientific discourse, and therefore offers a bridge between the philosophy and the science of consciousness: it offers an approach to conscious experience which, on the one hand, tries to account for the philosophically most important features of consciousness, whereas, on the other hand, does it in a way which smoothly fits into the everyday practice of scientific research.

Introduction

Humans are conscious beings and this fact permeates our whole life. Conscious experience is our access point to the world. What remains unconscious might trigger certain actions, nevertheless it does not contribute directly to how we see our environment. Literally, the content of our conscious experience is all there is for us. Moreover, the reach of consciousness spreads far beyond perception and knowledge —it features, for example, in our ethical judgments, and brings free will about. Understanding what consciousness is, and how it is connected to physical processes, thus, would be a key milestone along the way towards reconciling our subjective mental life with the material world sciences inform us about.

Consciousness presents the world in a certain way. There is something it is like to see, smell, and taste, for example, a fresh strawberry. These felt qualities determine the way a strawberry appears to us; they determine what a strawberry *is for us*. There is always a qualitative or phenomenal aspect characteristic of every conscious experience, and exactly these phenomenal qualities are what populate our subjective mental life. Therefore, a proper scientific account of consciousness should involve an explanation of what phenomenal qualities are and how they are related to physical processes.

However, according to the fundamental message of contemporary philosophy of mind, phenomenal qualities of conscious experience pose a serious problem for physical explanations. The heart of the problem is that the qualitative aspect of consciousness seems to *resist functionalisation*: it seems that phenomenal qualities cannot be adequately defined solely in terms of functional or causal roles (Chalmers, 1996). Even if the causal roles played by a particular conscious experience get identified, it remains a further question why that specific phenomenal quality is experienced when the causal roles in question are filled.

That is, there seems to be an *explanatory gap* between explaining the phenomenal qualities of conscious experience, on the one hand, and explaining physical structures and functional processes, on the other (Levine, 1983). This problem is especially pressing once one recognises that what physical explanations are typically able to account for are structures and functions. Phenomenal qualities, then, seem to be in principle unexplainable in physical—and hence scientific—terms.

This pessimistic conclusion has fundamentally affected the attitude of scientific approaches to consciousness. Instead of pursuing the original question—asking for an explanation of how consciousness is related to physical processes—scientists now mainly focus on finding what those physical processes are which co-occur with conscious experiences. Finding these so-called *neural correlates* became the primary goal of contemporary scientific approaches to consciousness (cf. e.g. Dehaene & Naccache, 2001).

Acknowledging that there is an explanatory gap between our conscious experiences and scientific knowledge, thus, results in a *further gap*, one between the philosophy and the science of consciousness. Philosophy and science have worked together for centuries throwing light on issues like the nature of life or matter. This alliance, however, seems to be at the verge of breakup when they turn towards the characteristics of our subjective mental life—whereas the philosophical approach aims at understanding the fundamental features of conscious experience, the scientific approach concentrates merely on pinpointing its neural correlates. This gap is quite prominent in contemporary interdisciplinary discourse. Scientists—since due to its functional un-analysability cannot adequately operationalise the qualitative aspect of consciousness—try to re-define conscious experience in cognitive and neural terms. Philosophers, at the same time, relentlessly argue that though such scientific theories might be theories of cognitive access, for example, but are definitely not theories of consciousness itself, since they cannot account for the phenomenal qualities of our conscious experiences (Tononi, 2004; Dehaene, et al., 2006; Lamme, 2006).

The fundamental aim of this dissertation is to bridge this gap between the philosophy and science of consciousness. I try to show that science is still able to resolve the mystery related to conscious experience.

However, bridging the gap between the philosophy and science of consciousness does not amount to bridging the explanatory gap itself. On the contrary: acknowledging that there is an explanatory gap is a fundamental tenet, a starting point of my approach. Phenomenal qualities might well be functionally un-analysable, nevertheless, the very fact that they resist functionalisation, and that they give rise to an explanatory gap are distinguishing features of conscious experience. Accounting for these—philosophically most important—characteristics of consciousness in purely scientific terms is the very purpose of my doctoral dissertation.

After setting the stage by discussing how the doctrine of physicalism might best be formulated (cf. Chapter 1), how the so-called epistemic gap between our phenomenal and physical knowledge gets established (cf. Chapter 2), and how Phenomenal Concept Strategy, the received view in contemporary physicalist literature, explains the presence of the epistemic gap (cf. Chapter 3), I present the main line of thought of the dissertation. I shift the focus from conceptual features to features of sensory/perceptual representations and formulate my own account, the so-called Monadic Marker Account as an alternative to Phenomenal Concept Strategy. The Monadic Marker Account pursues the purpose of bridging the gap between the philosophy and science of consciousness via five consecutive steps

First, I investigate the structure present in conscious experiences. I argue that typically, experiences are complex—they have discernible structure with constituent parts, which, in themselves, could occur as contents of standalone experiences. Some experiences, however, are simple, without any such structure—they have no discernible parts that could be experienced on their own. Having made these observations, I formulate and provide support for the central tenet of the first step of the dissertation's main line of thought. I argue that the phenomenal quality of having a complex experience is jointly determined by the discernible structure of the constituent parts, and the phenomenal aspect of the (ultimately) simple experiences of these constituents. (Cf. Chapter 4: §4.1.)

As the second step, I turn towards the cognitive-representational domain, which, in accordance with current scientific approaches, underlies conscious experiences. It is generally accepted by cognitive and neural theories that the representations involved in the processing of perceptual stimuli are organised into a hierarchy, such that representations at each level stand for parts of the features representations at the next higher level stand for (Biederman, 1987; Hayworth & Biederman, 2006; Körding & Wolpert, 2006; Mamassian, 2006; Yuille & Kersten, 2006; Kouider, et al., 2010). I argue that if we add a further assumption to this picture than it becomes possible to understand how conscious experiences are related to the cognitive-representational system. The assumption in question claims that central processes (working memory, global workspace) have access only to a range of the levels of this hierarchy. By accessing multiple levels within this range simultaneously central processes are able to extract structural information related to the arrangement of the parts constituting represented features, which makes representations at higher levels of the accessible range structured for central processes. Contrary to this, even if within the entire hierarchy there are lower-level representations standing for constituents of the features representations at the *earliest stage of the accessible range* stand for, central processes have no access to them. Therefore, representations at the earliest stage of the range in question are *monadic for central processes*—central processes cannot extract structural information about the features they stand for. These monadic representations are the most basic meaningful units, which carry information for the rest of the system; none of their properties can independently be interpreted by processes in the central system. (Cf. Chapter 4: §4.2, §4.3.1, §4.3.2.)

As the third step, I point out that there are essential similarities between the characteristics of the phenomenal realm uncovered in the first step and the features of the cognitive domain discussed in the second step. First, subjects are able to discern structure present in their complex conscious experiences; similarly, central processes are able to extract structural information from perceptual representations within the range of the hierarchy of representations central processes have access to. Second, subjects are unable to discern structure in their simple conscious experiences; similarly central processes are unable to extract structural information from monadic representations. Third, the phenomenal quality of complex experiences is determined by the phenomenal quality of simple experiences plus the structure characterising the arrangement of the constituent parts discernible for the subjects; similarly, the particular way higher-level representations are structured for central processes is determined by the lower-level (ultimately monadic) representations, plus the structural information extractable by central processes. That is, the role simple experiences play in the phenomenal domain is the very same role monadic representations play in the cognitive domain. On the basis of this similarity, I propose the following identity claim: the *phenomenal qualities of simple conscious experiences are identical with how monadic representations present themselves for central processes of a cognitive system*. (Cf. Chapter 4: §4.3.3.)

In the fourth step, I illustrate how the Monadic Marker Account works. I provide support for the identity statement proposed by showing that if one accepts the identity claim, then one will become able to account for why phenomenal qualities resist functionalisation and why they give rise to the explanatory gap solely in terms of the features of the cognitive-representational system. Monadic representations do not map the structure, only indicate the presence of the object they stand for. Any kind of representation with whatever features is apt for playing the role of a monadic representation in so far as the rest of the system treats it as a monadic whole. This feature, I argue, is able to account for the fact that phenomenal qualities resist functionalisation. Similarly, the explanatory gap necessarily arises in conscious cognitive systems deploying monadic representations. The source of the explanatory gap is that there is no *a priori* connection between scientific knowledge and how a cognitive system itself interprets monadic representations. On the one hand, access to how the system interprets monadic representations is privileged—it is restricted to the system itself,—whereas, on the other hand, the system lacks access to any of those features scientific investigations can provide information about. (Cf. Chapter 5.)

Ultimately, I am arguing for an identity theory. However, within philosophy of mind such identity claims have been severely criticised on the grounds that (a) they are so-called *brute* (i.e. unexplained) identities, hence we do not have reasons to believe in them, and that (b) they are unique, i.e. unlike any other typical identity claim occurring in scientific explanations (Chalmers & Jackson, 2001; Chalmers, 2010a). The final, fifth step of my dissertation answers these challenges. I show that, contra (a), there are reasons for believing in the particular identity claim I propose, and, contra (b), standard scientific explanations always deploy identity claims playing the same role the one proposed here plays. I argue that the fact that if one accepts the identity claim in question then explaining features of consciousness in scientific terms becomes possible is sufficient for believing in the identity claim, especially if one considers usual cases of scientific explanations (cf. Chapter 5: §5.3, and Chapter 7: §7.3). Moreover, I argue that identities play a special role in reductive explanation, which is significantly different from what is implied either by the transparent version of reductive explanation (Levine, 1993; Chalmers and Jackson, 2001; Kim, 2005), or by the inference to the best explanation based approach (Block and Stalnaker, 1999; McLaughlin, 2010). I show that typical scientific explanations deploy ‘prior identities’, i.e. identity claims, which themselves are unexplained but are nevertheless necessary for mapping the features to be explained onto the features the explanation relies on. The identity claim connecting the phenomenal character of simple experiences to monadic perceptual representations fills the very same role: it is an unexplained explainer, deployed in order to uncover similarities between the phenomenal and the cognitive domains. Once these identities are accepted, similarities can be revealed, and standard scientific explanations—of, for example, why phenomenal qualities resist functionalisation and why they give rise to the explanatory gap—can be formulated. (Cf. Chapter 6 and Chapter 7.)

Part 1

Setting the Stage

Chapter 1:
Physicalism

1.1 The Doctrine of Physicalism

The broad framework of this doctoral dissertation is provided by a debate over whether the doctrine of physicalism is true or not. Physicalism, roughly speaking, is the view that everything is physical, or as it is often put, the view that there is nothing over and above the physical. The term originates from Otto Neurath (Neurath, 1931a, 1931b); however, what Neurath meant by this expression was quite different from how it is understood in contemporary literature. According to Neurath, the term physicalism stood for the view that for every statement there is a physical statement equivalent in meaning with it (cf. Stoljar, 2009a). Contrary to this, physicalism today is usually understood as a metaphysical claim about the nature of the world rather than a linguistic claim about statements.

The fundamental tenet of the doctrine is often captured metaphorically. David Lewis, for example, says that according to physicalism, by copying the physical realm (all the physical properties) one copies all the facts of the world (Lewis, 1983b). Or as Tim Crane (1991, 2001b) explains the view: once God had created all the physical properties and laws God's work was done—all other properties came for free (cf. also Kripke, 1980; Owens, 1992). What these metaphors try to capture is a strong dependence relation between the physical facts and all other facts. Once the physical facts are fixed, they necessitate all other facts.

Intuitively, what counts as physical must be related to the science of physics. If so, however, then there are lots of facts—e.g. chemical, biological, psychological, social facts—which, at least on the face of it, seem to be non-physical. There are chemical forces, biological functions, mental states, social relations, etc. which fall outside the scope of physics. From this perspective, what physicalism tells us is that these prima facie non-physical phenomena are in fact physically acceptable (cf. Wilson, 2005,

2010) in the sense that they are necessitated by the physical realm. According to physicalism, there are no facts about chemical forces, biological functions, mental states or social relations (etc.) over and above the physical facts.

In a nutshell, this is the doctrine of physicalism. However, in order to have a thorough understanding of the thesis and its consequences, one needs to clarify what exactly is to be taken as physical and how to understand the ‘nothing over and above’ locution properly. §1.3-§1.5 and §1.6 deal with these issues in reverse order.

Before going into these details, though, it might be useful to close this introductory section with a terminological point. In contemporary literature the terms ‘physicalism’ and ‘materialism’ are often used interchangeably. As a matter of fact, however, the concept of materialism is a much older one than that of physicalism. According to the classical understanding, materialism is the view that everything is made up of matter. Material entities were typically characterised as being extended, impenetrable etc. Later the classical notion of materialism evolved into the modern notion of physicalism (cf. Wilson, 2006). The transition was influenced by those scientific discoveries which pointed out that the ultimate building blocks of nature like e.g. force-fields shared very few features of material entities taken in the classical sense. Thus instead of relying on the classical understanding of matter the doctrine of physicalism relies on the notion of physical which is supposed to stand for a scientifically informed characterisation of the fundamental entities building up nature. However, those who use the term ‘materialism’ in contemporary debates do not evoke its classical understanding—what they (e.g. Chalmers, 1996; Levine, 2001; Papineau, 2002) have in mind is the very same doctrine as the one discussed by those employing the term ‘physicalism’ (e.g. Loar, 1990; Crane, 2001b; Loewer, 2001); i.e. the doctrine relying on the notion of physical rather than that of matter. Nevertheless, since the notion of materialism is more or less still loaded with its classical interpretation, throughout this dissertation I shall follow those who refer to the doctrine in question as physicalism.

1.2 The Causal Argument for Physicalism

Once we have an intuitive grasp on the concept of physicalism, the question what reasons its proponents have to hold the view arises. The current section deals with this problem—it introduces the case for physicalism, i.e. the main argument supporting the view. In this section my aim is only to briefly introduce the argument itself. In contemporary literature both the validity and the soundness of the argument is hotly debated (e.g. Yablo, 1992; Sturgeon, 1998; List & Menzies, 2009; Menzies & List, 2010; Raatikainen, 2010). Nevertheless, for our present purposes many of the details of these debates are unimportant.

The line of thought which is most widely referred to as the main argument for physicalism has many names, e.g. the causal argument (Papineau, 2002), the argument from causal closure (Stoljar, 2009a), or the causal exclusion argument (Kim, 1993b, 1998, 2005). This argument stems from the observation that prima facie non-physical (e.g. chemical, biological, mental, social, etc.) entities causally interact with physical entities. However, all physical events seem to have sufficient physical causes. So unless all those physical effects which have sufficient prima facie non-physical causes are systematically caused twice over, prima facie non-physical causes must be identical with physical causes. To put it more formally, the general version of the argument runs through the following premises.

P1: Prima facie non-physical causes are sufficient for certain physical effects.

P2: All physical effects have sufficient physical causes.

P3: The physical effects of prima facie non-physical causes are not all overdetermined.

The first premise tells us that certain physical effects have efficient prima facie non-physical causes. The second premise tells us that all physical effects—and hence those having prima facie non-physical causes—have efficient physical causes. The third premise tells us that those physical effects, which are claimed to have two distinct causes (a physical and a prima facie non-physical) are, in fact, not caused twice over. From these premises, the causal argument says, it follows that prima facie non-physical causes are, in fact, physical causes. That is, the causal argument motivates the doctrine of physicalism (claiming that there is nothing over and above the physical) by showing that those causes which prima facie appear to be non-physical, turn out to be physical after all.

1.2.1 Prima facie non-physical causes having physical effects

Now let's consider each premise in more detail. The first premise captures the very observation motivating the argument itself: that there are certain prima facie non-physical events causing physical events. For example, the event that I feel the urge to reduce my thirst (which is a mental, i.e. prima facie non-physical event)—together with my beliefs that there is water inside that bottle in front of me and if I lift it to my mouth I can satisfy that desire—results in my hand grabbing the bottle and lifting it to my mouth (which is a physical event). That is, it seems to be an observational fact that mental events cause physical events (and similarly with other prima facie non-physical, e.g. biological, social, etc. events).

Nevertheless, the causal argument is sometimes called into question via debating its first premise. Approaches following this line of thought do not accept that mental (or any other prima facie non-physical) events are causes of physical events. For example, Gottfried Wilhelm Leibniz suggested that there were no connections between the mental and the physical realms. According to his view, neither mental events cause physical events, nor vice versa. Instead, Leibniz argued, the apparent correlation between mental and physical events (e.g. that the occurrences of the desire to reduce thirst are typically followed by lifting some bottle) are to be explained by relying on a so-called pre-established harmony: God has arranged the physical and the mental realms so that they 'run parallel with each other' thereby guaranteeing these correlations (Leibniz, 1898).

Another way to deny that prima facie non-physical events cause physical events is to subscribe to the view called epiphenomenalism. According to epiphenomenalism, although prima facie non-physical (e.g. mental) events are caused by physical events, they themselves do not cause any physical or other events—they are causal dead-ends. For example, if one observes the shadows casted by two colliding billiard balls one might want to conclude that there is an incoming shadow pushing away the other shadow. Contrary to this, however, there is no causal connection between the two shadows: all the movement one observes is the result of the movement of the real billiard balls (and the light rays coming from a certain source).[1]

1.2.2 The causal completeness of the physical

The second premise captures the idea that the physical realm is causally closed—one, who is looking for sufficient causes of physical effects will never need to leave the physical realm to find them. There is a causal chain leading to any physical effect consisting of solely physical causes. This is the so-called *causal completeness* or *causal closure* thesis.

Note that the physical realm is quite unique with regard to this feature: no other realms are complete in this sense. For example, there are economical effects which have no sufficient economical causes, rather they are brought about by certain sociological or psychological causes (cf. e.g. the role of rational agents and their psychological needs in economical theorising). Similarly, the realm of sociology is incomplete as well: there are certain non-sociological (e.g. physical—say, a tsunami) events which are sufficient causes of certain sociological effects. Or consider the classical biological example: the occurrence of a certain mutation is a biological effect, the sufficient cause of which is typically a physical event, e.g. a high-energy

[1] Epiphenomenalism is most often criticised on the basis that it presupposes causal dead-ends. The problem with this commitment of epiphenomenalism is that it goes against how we typically account for natural phenomena. As Papineau puts it: "So if epiphenomenalism were true, then the relation between the mental and the physical would be like nothing else in nature, since science recognises no other examples of 'causal danglers', ontologically independent states with causes but no effects." (Papineau, 2002, p. 23). Also cf. Smart (1959), Jackson (1982), Chalmers (1996), Kim (1998, 2005).

photon hitting the DNA. That is, there are causal chains ‘leading out’ of the realms of economics, sociology, psychology, biology, etc. It is only the realm of physics which seems to be general or broad enough so that all physical effects have sufficient physical causes. Whether the physical is really causally complete, is, of course, an empirical issue. However, for example, David Papineau (2000) argues in length that the thesis has profoundly been established over the last 150 years.[2]

The prototypical view straightforwardly denying the causal completeness of the physical is Descrates’ interactionalist dualism. Descrates’ view claims that there are sui generis mental causes which interact with the physical realm and bring about physical effects which themselves have no sufficient physical causes (Descartes, 1984). That is, the physical realm, according to interactionist dualism is not closed—those physical events which are the effects of mental causes would have not occurred had mental events not caused them.[3]

Similarly, classical emergentism (cf. e.g. Alexander, 1920; Morgan, 1923; Broad, 1925) claims that there are non-physical (so-called emergent) properties instantiated by certain compound objects which bestow novel causal powers upon the objects instantiating them. These novel causal powers affect the physical base out of which the special non-physical properties in question emerge. This so-called downward causation makes emergent properties causally efficacious and autonomous, which in turn entails that the instantiation of emergent properties violates the causal completeness of the physical.[4]

More recently, Peter Menzies and his colleagues have come forth with a proposal claiming that under the so-called interventionist account of causation (Woodward,

[2] See also Spurrett and Papineau (1999) and especially David Spurrett’s doctoral dissertation (1999) for a detailed analysis of the causal completeness of the physical.

[3] According to Descartes, the mental interacts with the physical via the *pineal gland*, where it alters the direction of bodily movements. (Cf. Papineau, 2000)

[4] See §1.4 for a detailed discussion of emergentism. §1.4.2 explores the problem of emergent-physical causation in depth.

1997, 2003) the mental excludes the physical as the cause of certain physical effects. That is, the Menzies-proposal tries to show that although the causal completeness of the physical might be an attractive thesis under some (e.g. productive) accounts of causation, it seems to be doubtful if one subscribes to alternative accounts of causation (List & Menzies, 2009; Menzies & List, 2010; Raatikainen, 2010).

1.2.3 No overdetermination

The third premise captures the idea that systematic overdetermination is unlikely. If prima facie non-physical events do cause physical events (in accordance with P1), and if these physical events also have sufficient physical causes (in accordance with P2), and if the physical and prima facie non-physical events are distinct and independent of each other, then the physical events in question will be caused twice over by two independent causes.

Overdetermination in itself, of course, is not something impossible. For instance, if two assassins, independently of each other, shot the same victim in such a way that the two bullets penetrated the victim’s heart at the same time, then the victim’s death would be overdetermined by the two shots. The two shots would be two independent causes bringing about the same effect (the death of the given victim at the given time). The victim would have died even if one of the assassins had missed her shot.

A case like this, however, is a case of pure coincidence. Had one of the shots hit the heart of the victim only a second earlier (and supposing that death strikes instantaneously after a heart-shot) the victim would have already been dead when got hit by the second shot. This situation would no longer be a case of overdetermination, but rather a case of causal preemption—the first shot would have causally preempted the second shot: it would have caused the effect in itself (single cause, single effect) leaving no room for the second shot to display its causal potency.

P1 and P2 together suggest that whenever a prima facie non-physical event causes a physical event, there is always a physical event as well, kicking in in exactly the same moment, and causing the very same effect. The no overdetermination thesis as captured by the third premise of the causal argument for physicalism expresses that this kind of *systematic* overdetermination is highly unlikely. Moreover, systematic overdetermination calls for an underlying mechanism, ensuring that a certain class of physical events always has two (or more) independent and individually sufficient causes. No evidence supports that such a mechanism exists. The no systematic overdetermination thesis even coincides with our intuitions with regard to causation: we implicitly presuppose that an effect has a single sufficient cause. For example, the police stops its investigation when it finds the perpetrator; it does not continue to look for a second, independent one. So it seems reasonable to appeal to the third premise as well (cf. Papineau, 2002, pp. 26-28; Kim, 2005, pp. 46-52).[5]

1.3 Understanding 'Nothing Over and Above'

Now that we have a rough sketch of the main argument supporting the doctrine of physicalism at hand, it is time to dive into the details of how to understand the view properly. It is a common coin that according to physicalism, prima facie non-physical facts are nothing over and above the physical facts. However, the exact nature of this relation between the physical and the prima facie non-physical is controversial.

1.3.1 Eliminative physicalism

The most straightforward interpretation of the 'nothing over and above' clause is elimination: claiming that all there is is, literally, only the physical. Eliminative physicalism challenges the existence of those entities which get identified via apparently non-physical terms. The classical example of this view is the elimination

[5] For an argument biting the bullet, and claiming that mental causation is a genuine case of overdetermination see for example Mellor's (1995) 'belts and braces' view. Note that Kim's version of the causal arguments is targeting the so-called non-reductive version of physicalism (cf. §1.3.3). In this case, though, the no overdetermination thesis is not that straightforward, since the two proclaimed causes (say, a mental cause and its subvenient physical base) are not ontologically independent (cf. e.g. Yablo, 1992; Carey, 2011).

of phlogiston from the scientific theory of combustion. Briefly, phlogiston theory states that all flammable materials contain a special ingredient, called phlogiston, which is released when the material in question burns. However, subsequent scientific discoveries revealed serious flaws in the theory, and finally, the phlogiston theory had been replaced by the oxygen theory of combustion. As it turns out, the role phlogiston played in the phlogiston theory is almost exactly the opposite of the role oxygen plays in the oxygen theory. The conclusion is that there is no such thing in nature as phlogiston—what there is is oxygen instead. To put it in another way: the term ‘phlogiston’ does not refer at all. That is, eliminative physicalism denies the existence of the entity eliminated. It claims that such an entity is a postulation of a seriously flawed, fundamentally mistaken theory, and as a theoretical term is non-referring: there is nothing in nature which is picked out by the term. This is a strong claim rendering certain terms empty.[6]

Historically a much weaker claim is also associated with eliminative materialism. The difference between the two interpretations is that the weaker understanding of eliminativism acknowledges that the terms of the ‘mistaken’ theory refer, but only inaccurately—there are better, more accurate, ways to pick out their referents, and these better ways will eventually take over the place of the mistaken theory. Consider, for example, the relation between Newtonian mechanics and special relativity—especially their use of the term ‘mass’. Unlike in the case of phlogiston, the relativistic theory succeeding Newtonian mechanics does not dispense with the term ‘mass’, but rather claims that its definition requires some refinement: what the Newtonian term ‘mass’ picks out is only a limiting case of what the relativistic term ‘mass’ stands for. Note how different the strong claim is from the weak claim. The strong claim entails that certain prima facie non-physical properties do not exist, whereas the weaker claim is compatible with the view that prima facie non-physical

[6] In contemporary philosophy of mind, eliminative physicalism stands for the view that common-sense understanding of the mental is mistaken and our ordinary notions of mental states are empty (cf. P. M. Churchland, 1981; P. S. Churchland, 1986).

properties do exist, and are *indeed* physical properties: prima facie non-physical terms stand, though inaccurately for physical properties (cf. Ramsey, 2011).[7]

In the course of this dissertation I follow the distinction introduced by Savitt (1974) between eliminativism and reductionism. According to this understanding, eliminativism is ontologically radical, whereas reductionism is ontologically conservative. That is, the fundamental tenet of eliminativism is best characterised by the strong claim above, i.e. by the claim that the terms to be eliminated designate nothing in nature. The weaker claim, on the other hand, characterising the terms to be eliminated as inaccurate designators of physical properties, is closer to the view called reductionism.

1.3.2 Reductive physicalism

Considering only those possibilities which do not deny the existence of prima facie non-physical properties, the most straightforward interpretation of the 'nothing over and above' clause is identity: prima facie non-physical facts depend on physical facts because each and every prima facie non-physical property is identical with a physical property (or set of properties). According to this interpretation, then, those properties which on the face of it appear to be non-physical are in fact physical. That is, those properties which are initially identified as e.g. chemical, biological, psychological etc. properties are claimed to be physical properties.[8]

The most famous version of this interpretation of physicalism is the so-called identity theory of the mind (Place, 1956; Feigl, 1958; Smart, 1959). The identity theory is a

[7] Rorty (1965) is a good example of confusing the stronger claim with the weaker one. See Conman (1968), Lycan and Pappas (1972), and Savitt (1974) for clarifying the distinction between the two claims.

[8] Of course, a lot turns on how we define what it is to be physical. See §1.6 for more on this question. For our present purposes, it is enough to note that under a theory-based definition of physical (saying that something is physical if it is picked out by a term of physics) one might feel tempted to object against the identification of prima facie non-physical properties with physical properties on the basis of the fact that prima facie non-physical properties are picked out by certain terms of chemistry, biology, etc. Notice, however, that the criterion that something is physical if it is picked out by a term of physics is only a sufficient condition—it is compatible with the same thing being picked out by a term of physics and a term of, say, chemistry.

particular thesis about mental states. It claims that mental states and events are not just correlated with but identical to brain processes. That is, say, having an experience of seeing something red is identical with having brain process XYZ. At first sight, this might be a quite contra-intuitive claim, for perceptual experiences seem to be very different from brain processes. For example, there is nothing red in the brain (brain haemorrhage put aside), and similarly the experience of seeing something red is hard to be localised in space, or attributed with extension. The identity theory answers this problem by emphasising that *having* an experience of red is not itself red.[9] The identity theory identifies the property of having an experience of red with the property of having brain process XYZ. Put in another way, the terms 'having an experience of red' and 'having brain process XYZ' have the same referents. They differ in meaning, nevertheless they pick out the same thing—just like the terms 'Morning Star' and 'Evening Star' have different meanings, still they both refer to planet Venus (cf. Smart, 2008). That is, the point identity theory makes is this: instead of there being two distinct entities (a mental event and physical process) there is only one thing which can be picked out in two quite distinct ways.[10]

Note that contrary to what some might think, the identity theory does not eliminate mental phenomena. Reading the fundamental claim of the identity theory as saying 'all there *really* is is *just* the physical' is misreading it. Identifying two objects eliminates neither of them. Claiming that Bob Dylan is Robert Allen Zimmerman does not imply the non-existence of either Bob Dylan or Robert Allen Zimmerman. On the contrary, an identity ensures that both Bob Dylan and Robert Allen Zimmerman exists—and it ads the further claim that they happen to be the same person. True, an identity statement eliminates the number of entities that would exist if the identity statement were false but does not eliminate any of the named objects. For example, someone, who without knowing their identity would have thought that Bob Dylan and Robert Allen Zimmerman were two separate persons, would now

[9] Cf. what U. T. Place calls 'the phenomenological fallacy' (Place, 1956). See also in Smart (1980, pp. 111-112).

[10] See §6.3.2 for more details.

know that there is only one person. However, she would not conclude that any of her earlier thoughts like ‘Bob Dylan exists’ or ‘Robert Allen Zimmerman exists’ was wrong.

Similarly with scientific identity statements like water is identical to H_2O. The claim here is that the entity which under certain conditions (say, in everyday life, observing it, for example, in a lake) is called ‘water’ is the same entity which under different conditions (in scientific practice, observing it, for example under an electron-microscope) is called ‘H_2O’. Identity statements claim that there is one entity that can be picked out in two different ways. That is, according to identity statements both of the terms in question refer. Recall that this is exactly what is denied by eliminativism. For example, when one argues for eliminating phlogiston what one implies is that the term phlogiston does not refer—there is nothing in the world that is picked out by the term.[11]

The identity theory is often considered as a version of reductive physicalism. However, the relation between the identity theory and reductionism needs clarification as the concept of reduction has multiple connotations. According to the so-called functional model of reduction (Kim, 1998, 1999, 2005)—which is the received view in contemporary philosophy of mind (cf. e.g. Levine, 1993; Chalmers, 1996; Van Gulick, 2001)—a property gets reduced to another property via a process consisting of the following steps. First the property to be reduced is redefined in functional terms, i.e. in terms of a functional role. Second, the property in the reducing base which actually plays the given functional role is identified. And third, this then gets supplemented by a theory explaining how the property of the reducing base is able to perform the functional role in question. Thus a property gets reduced to another property.[12] Jaegwon Kim, for example, uses functional reduction to show

[11] Identity statement, especially the similarity of mind-brain identities to classical cases of scientific identities or to identity claims connecting proper names is of special interest in this dissertation. The issue emerges over and over again throughout the dissertation and culminates in chapters 6 and 7, which are entirely devoted to this topic.

[12] See §6.2.3 for more details on functional reduction.

that mental properties reduce to the physical base properties actually realising them. The conclusion is often formulated as an identity claim: mental property M is identical to physical property P. This would, then, suggest that reduction amounts to formulating an identity statement. Note, however, that reduction cannot be synonymous with identification: reduction is asymmetrical whereas identification is symmetrical. While mental property M is reduced to physical property P in the above model of functional reduction, the opposite is not true: property P is not reduced to property M. Contrary to this, identities are symmetrical: if property P is identical with property M then property M is identical with property P as well.

According to Tim Crane, this distinction between reduction and identification points out that there are two different ways in which one can use the notion of reduction (Crane, 2001a). One way of understanding reduction is explanatory: a phenomenon is reduced in this sense if by this process it is "made more comprehensible or intelligible" (Crane, 2001a, p. 54). The explanatory aspect of reduction is especially prevailing in so-called mechanistic reductions which are trendy extensions of the Kimian functional model of reduction in contemporary philosophy of science (Bechtel, 2007, 2008). Mechanistic reductions explain why a certain system is able to do what it does via decomposing it into a mechanism (an organised structure) of its working parts. Mechanistic reductions show, for example, that a given whole S performs a certain task because (1) it is identical with the organised structure of entities X_1, X_2 etc., and (2) these entities when organised spatially and temporally in the right sort of way together perform the task in question.[13] This explanatory connotation of the general notion of reduction is the main reason why reduction is not analogous with identification—the 'making more intelligible' relation is not symmetrical: whereas, for instance, understanding the organisation of X_1, X_2, etc. gives some insight into the workings of S, the opposite is not true. Explanatory reduction is typically a relation between theoretical terms of different theories, and plays a crucial role in inter-theoretical reductions in philosophy of science.

[13] See §7.2.1 for more details on mechanistic explanation.

The other understanding of reduction is reduction in a strict ontological sense: a relation between entities certain terms pick out. Claiming that mental property M is reduced to physical property P in this ontological sense expresses that the referent of the term 'mental property M' is the same as the referent of the term 'physical property P'. That is, ontologically reducing a certain property to another one amounts to identifying the two properties. Thus, it is the ontological aspect of reduction which is analogous to identification, and therefore relevant in metaphysical contexts. That is, if one is concerned with versions of physicalism—as we are here—then, given that physicalism is a metaphysical thesis, reductive physicalism is correctly understood as an ontological thesis claiming that prima facie non-physical properties are identical with physical properties.[14]

1.3.3 Non-reductive physicalism

Not more than a decade after the heyday of the identity theory in philosophy of mind, and only a few years after that the defining piece on reduction in general philosophy of science had been published (E. Nagel, 1961), Hilary Putnam proposed an argument which almost immediately led to the demise of reductive physicalism (Putnam, 1967). Putnam pointed out that the idea behind identity theory was extremely implausible. For, thus Putnam's line of thought goes, the identity theory requires that the physical properties corresponding to a mental state must be the same in all those creatures which can share the same mental state. However, biologically quite distinct beings like humans and, say, octopi produce, for example, similar pain avoidance behaviour and therefore likely possess the mental state of feeling pain. Since their biological organisation (including evolutionary history) is quite different, arguing that nevertheless the physical property corresponding to feeling pain in humans and in octopi is the very same property would be an extremely strong claim. Note, that Putnam's argument does not show that the identity theory must be wrong,

[14] Chapters 6 and 7 present a more detailed analysis of the concept of reduction in general, the role identities play in reductions, and the relation between identity statements and explanation in particular.

it only claims that the identity theory is empirically unlikely.[15] Putnam's alternative suggestion is that mental properties, rather than being identical to physical properties are realised by them. Since, according to Putnam's claim, the very same mental state can be realised by many different physical states, mental states are multiply realised.[16]

Putnam's point is that the identity theory cannot incorporate cases of multiple realisability, and therefore must be abandoned in favour of a view that can. The view, which, according to Putnam, is best suited for this task (i.e. dealing with cases of multiple realisability) is functionalism. That is, Putnam argues for functionalism, as opposed to identity theory, on the grounds that in the light of multiple realisability, the former is more plausible than the latter. Functionalism, roughly, is the view that certain properties are identified via their functions, i.e. the roles they fill, the causal connections they have to other properties. In our present context, functionalism can be put as a view claiming that those properties which, on the face of it, appear to be non-physical, are in fact functional properties, strictly distinct from (i.e. not identical with) physical properties.[17]

[15] However, it is possible to strengthen Putnam's original argument by 'going modal', and claiming that the pertinent identities—if true at all—are necessarily true. In this case, then, the mere possibility of multiple realisability falsifies the identity theory.

[16] Soon after Putnam's seminal paper, Jerry Fodor extended the multiple realisability argument to philosophy of science in general (Fodor, 1974). Fodor argues against the classical Nagelian view of inter-theoretical reduction, according to which all special science kinds are—with nomological necessity—coextensive with physical kinds (see §6.2.1 for more on Nagelian reduction). Fodor claims that special sciences are very much in the business of formulating interesting generalisations about events whose physical descriptions have nothing in common. In the Fodorian picture special science kinds do not reduce to lower-level special science kinds (and ultimately to physical kinds) in the standard Nagelian sense. Fodor's proposal represents special science kinds as being multiply realised by the lower-level kinds: the antecedent and the consequence figuring in a special-science law are each connected with a disjunction of predicates in the lower-level science. The kinds of the higher-level science are realised by very different kinds in the lower-level science. This contradicts Nagelian reductive attempts, Fodor argues, since though each of the predicates at the lower level—one-by-one—are connected by a proper law of the lower-level science, their disjunctive generalisations, due to the variety of the realisers, are not law-like.

[17] Note that some proponents of functionalism do not follow this understanding (which is called role functionalism). Jaegwon Kim, for example, argues that functionally defined properties are identical with their realisers (hence, this view is called realiser functionalism), and multiple realisability based objections can be overcome by evoking species-specific reductions (e.g. Kim, 1998, 2005). Cf. functional reduction as in §1.3.2 and §6.2.3.

Those versions of physicalism, which, being persuaded by the multiple realisability argument, reject identities as a viable interpretation of the 'nothing over and above' locution are in general called versions of *non-reductive physicalism*.[18] Non-reductive physicalism, thus, holds that there are properties distinct from physical properties which nevertheless depend on physical properties. The standard way of capturing this dependence relation—i.e. the most general interpretation of being nothing over and above the physical—is usually formulated via the notion of supervenience. Supervenience is a relation between two sets of properties A and B. A properties supervene on B properties if and only if there can be no changes in A properties without there being changes in B properties. Or to put it in another way: if A properties supervene on B properties then whenever any two objects instantiating both A and B properties are indistinguishable with respect to their B properties they must also be indistinguishable with respect to their A properties (cf. Davidson, 1970, p. 98).[19]

If one fills this into the slogan of physicalism (there is nothing over and above the physical) what one gets is this: prima facie non-physical properties supervene on physical properties. This way of putting physicalism is sometimes called *supervenience physicalism* (Stoljar, 2009a). Being committed to the supervenience thesis, however, is not sufficient for defining physicalism properly. As it has been argued by many (e.g. Horgan, 1993; Kim, 1999; Crane, 2001b) a pure supervenience-based account of physicalism fails to distinguish itself from emergentism. Emergentism[20] is usually understood as physicalism's best traditional

[18] See, however, Davidson (1970) for a view not falling into this category.

[19] If the definition quantifies over objects in the same world, the resulting version of supervenience is so-called weak supervenience. Similarly, if the definition quantifies over objects in all possible worlds, the result is strong supervenience. And finally, if the definition quantifies over entire possible worlds (talking about two possible worlds which if indistinguishable with respect to B properties are also indistinguishable with respect to A properties) then the resulting version is global supervenience. Weak supervenience is generally considered to be too weak to cover the needs of physicalism (since it is compatible with, for example, Leibniz' pre-established harmony). It is a common coin that strong supervenience entails global supervenience, however, whether global supervenience entails strong supervenience is hotly debated (cf. Kallestrup, 2011).

[20] At least its ontological version—see Footnote 21 below.

rival (Wilson, 2005). It is a version of a combination of substance-monism and property dualism holding that there are emergent properties which are genuinely novel in the sense that they are not consequences of physical properties, and still they are not entirely independent of them (cf. e.g. Broad, 1925; McLaughlin, 1992). According to the classical interpretation (based on Broad, 1925), emergents are: (E1) *basic*—they are neither identical with, nor derivative (not even in principle) from physical properties; (E2) *genuinely causal*—they bestow new causal powers on the particulars instantiating them; and (E3) *determined by physical properties*—they emerge only when an appropriate set of physical properties are instantiated.[21] Now the problem is this: the kind of determination emergentism is committed to is exactly what is meant by supervenience. That is, if we accepted a pure supervenience-based formulation of physicalism, then emergentism, due to (E3), would become a version of physicalism. This would pose a problem, since the ontological reading of (E1) and (E2) is in tension with any version of physicalism.

In the literature, this problem is usually solved by differentiating between the modal strength of the supervenience relation physicalism and emergentism endorse: physicalism claims that all prima facie non-physical properties supervene on physical properties with *metaphysical* necessity, whereas according to emergentism, there are properties (the emergent ones) which supervene on physical properties only with *nomological* necessity (van Cleve, 1990; McLaughlin, 1992; Kirk, 1996; Stoljar, 2000).

However, in an influential paper, Jessica Wilson (2005) questioned the viability of this distinction. She argues that interpreting emergentism as being committed only to nomological necessity is mistaken—given a certain plausible account of properties

[21] (E1)-(E3) allow at least two different readings: an epistemological and an ontological one. According to the epistemological reading, emergents are explanatorily basic—they are unpredictable from the knowledge of, and unexplainable in terms of the physical properties. The ontological reading says that emergents are metaphysically basic—they are over and above the physical properties in a metaphysical sense. (E1)-(E3) are compatible with Broad's view, and also with most contemporary accounts. Shrader (2009) posits a characterisation which is similar to my (E1)-(E3) ontologically understood, and argues that it is a set of necessary conditions for any account of ontological emergence.

and causal powers, emergentism, just like physicalism, deploys metaphysical necessity. In what follows, I will show that Wilson's argument doesn't go through. That is, I will argue that the 'nothing over and above' clause is best understood as metaphysical determination. However, in order to be able to run my argument at full force, first I need to clarify a few issues related to emergentism.

1.4 Interpreting 'Nothing Over and Above' as Metaphysical Determination

First, (E1) tells us that emergents are genuinely novel properties of a system. They are—as it is often put—properties of composite objects which neither are possessed by any of the constituents of the composites nor are the *resultants* (cf. Lewes, 1875) of the properties of the parts.[22] That is, emergent properties are not derivative[23] from the properties instantiated by the entities composing the composite object.

Second, (E2) expresses that emergents contribute new causal powers to the particular instantiating them. By instantiating an emergent property an individual gets in possession of a novel causal power which is not derivative from the causal powers bestowed by the properties instantiated by parts of the individual in question. That is, the behaviour (the effects) of a system instantiating emergent properties is different from the behaviour the (otherwise unchanged) system would have performed had it not instantiated emergent properties. To put is shortly: emergent properties are both causally efficacious and autonomous.

Third, what (E3) says is that though emergents are not resultants of the properties instantiated by the composing parts in the above sense, they are nevertheless

[22] Here I shall follow the British Emergentists' (Alexander, 1920; Morgan, 1923; Broad, 1925) understanding of emergence. On their account, emergent properties are always properties of composite objects. See, however, e.g. Batterman (2002) for an account of emergentism which denies the significance of part-whole relations.

[23] The term 'derivative' is intended to be neutral to the ontic/epistemic distinction here. The rest of the section clarifies how to give this neutral term a proper ontological reading.

determined by them in a certain way. As C. D. Broad formulated this determination relation:

> "the characteristic behaviour of a living body is completely determined by the nature and arrangement of the chemical compounds which compose it, in the sense that any whole which is composed of such compounds in such an arrangement will show vital behaviour and that nothing else will do so." (Broad, 1925, pp. 67-68)

In the following subsections my objective is to show how (E1), (E2), and (E3) might be best understood in order to get a coherent approach. In the course of this endeavour I shall distinguish two kinds of emergent laws and determine their nature. Let's start in reverse order.

1.4.1 Understanding (E3)

As the Broad quotation above tells us emergent properties are determined by physical properties[24] in the sense that whenever a certain set of physical properties is instantiated in a certain structure (i.e. whenever the components of a given composite form a specific order) the emergent property gets instantiated.[25]

Let's say that a composite object o is composed out of some components $c_1, c_2, \ldots, c_j$ in such a way that they form a structure $\underline{s}$ and instantiate physical properties $P_1, P_2, \ldots, P_k$. In this case the determination relation says that if object o instantiates emergent property E then every object o^* having some components $c_m, c_n, \ldots, c_r$ forming structure $\underline{s}$ and instantiating physical properties $P_1, P_2, \ldots, P_k$ will necessarily instantiate E. That is, according to the physical-emergent determination, it is the

[24] Specifying what counts as a *physical* property is a controversial issue (cf. Crane & Mellor, 1990). For present purposes take the term as referring to the properties instantiated by the parts of the composite object. See §1.6 for an extensive discussion of this issue.

[25] According to Broad—as the 'nothing else will do so' clause expresses,—the opposite is also true: the emergent property is instantiated only if a particular structure of certain physical properties is instantiated. However, in contemporary literature, it is often argued that emergent properties are multiply realised. (Cf. Bedau, 1997; Laughlin & Pines, 2000; Bedau, 2008) So, in order to keep the analysis presented here compatible with those views, which are committed to the multiple realisability of emergents, the physical-emergent determination shall be taken as a one-way determination relation.

instantiation of structure $\underline{s}$ of properties $P_1, P_2,\ldots, P_k$ (for short: $\langle P_1, P_2,\ldots, P_k \mid \underline{s} \rangle$) that gives rise to the instantiation of emergent property E.

The physical-emergent determination tells us that instantiating a structure $\underline{s}$ of physical properties $P_1, P_2,\ldots, P_k$ is a sufficient condition for instantiating emergent property E. This condition allows us to connect the particular structure of the physical properties to the emergent property itself by a connecting principle telling us a rule about when the emergent property in question gets instantiated: whenever the particular structure of the physical properties is instantiated. That is, these connecting principles characterise the instantiation of emergents—they define, so to speak, when emergent properties emerge out of physical properties. Connecting principles of this sort are what C.D. Broad calls *trans-ordinal laws*:[26]

> "A trans-ordinal law would be a statement of the irreducible fact that an aggregate composed of aggregates of the next lower order in such and such proportions and arrangements has such and such characteristic and non-deducible properties." (Broad, 1925, pp. 77-78)

Trans-ordinal laws connect emergent properties to the corresponding structures. In this sense, trans-ordinal laws assign emergent properties to the structures of the physical base property sets and determine the instantiation of the emergent properties in question. In accordance with the categoricalist[27] understanding of laws of nature, instantiations of certain structures of some physical properties determine the instantiation of an emergent property *because* there are trans-ordinal laws connecting them. That is, the occurrence of an emergent property is due to there being a trans-ordinal law connecting the emergent property to the particular physical base property

[26] Trans-ordinal laws are sometimes called emergent laws. Here I would like to avoid this terminology and reserve the term 'emergent law' for a broader set. It is a purpose of this and the following section to draw attention to the difference between two subsets of emergent laws.

[27] Categoricalists think that the roles properties play are inessential to their nature. Roughly, according to the categoricalist, in addition to properties, there are contingent N-relations (second order universals) connecting properties and telling them what to do. The fact that property P necessitates the instantiation of property Q is due to there being an N-relation connecting P and Q. Laws of nature are these contingent facts—e.g. N(P, Q). (Cf. Armstrong, 1983). To see how emergentism might be understood in a non-categoricalist (i.e. dispositional essentialist) framework, consult §1.5.5.

set. Emergent property E is instantiated by object o^* instantiating a structure $\underline{s}$ of properties $P_1, P_2, \ldots, P_k$ *in virtue of* a trans-ordinal law LTO ($\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle, E$).

Trans-ordinal laws tell us that whenever $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ gets instantiated E also gets instantiated. According to the interpretation of ontological emergence advocated here, neither the instantiation of $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ precedes the instantiation of E in time, nor vice versa. LTO-s are special laws of nature (partly) because they are not causal laws; what they express are synchronic, non-causal co-variations between the emergent properties and the corresponding structures of certain physical properties. Though, due to LTO-s, emergent properties are nomologically determined by physical properties, according to this interpretation, it is not the case that the instantiation of $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ *causes* the instantiation of E—rather $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ non-causally necessitates E.[28]

On the basis of all this, (E3) is to be understood in the following way. Physical properties determine emergents in the sense that instantiations of specific structures of certain physical properties non-causally necessitate the instantiations of corresponding emergent properties *in virtue of* some special laws of nature—trans-ordinal laws—connecting the instantiations of emergent properties to the instantiations of the requisite physical property-structures.

1.4.2 Understanding (E2)

(E2) says that emergent properties are causally efficacious and autonomous. This means that whenever an emergent property E gets instantiated by an object o it bestows certain novel causal powers upon object o. According to the ontological reading, these new causal powers are novel in the sense that they are metaphysically independent of the causal powers bestowed by the properties parts of object o instantiate. Those nomologically different possible worlds where $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$

[28] See §1.5.3 for more details about the causal vs. non-causal interpretations of trans-ordinal laws.

is instantiated with all the causal powers of $P_1, P_2, \ldots, P_k$ but the novel causal power of E is not manifested (say, because the relevant L_{TO} is not present and thus E does not appear) are causally different from the world where the novel causal power of E gets manifested as well.

The causal efficacy of emergent property E is manifested by causal connections between E and either other emergent properties or some physical properties.[29] Again, according to categoricalism, the causal link between emergent property E and physical property P_n is due to a *law of emergent causation* $\mathrm{L}_{\mathrm{EC}}(E, P_n)$ establishing a nomological connection between an emergent and a physical property. What $\mathrm{L}_{\mathrm{EC}}(E, P_n)$ tells us is a causal regularity: whenever emergent property E gets instantiated, it is followed by an instantiation of physical property P_n.[30] That is, though both L_{EC}-s and L_{TO}-s connect emergent properties with physical ones, they differ from each other in an important aspect—whereas L_{TO}-s express synchronic, non-causal co-variations, L_{EC}-s express causal regularities.

In the literature, the distinction between L_{EC}-s and L_{TO}-s is often overlooked. For example, O'Connor and Wong clearly do *not* make this distinction when they say that "[the] newness of [an emergent] property […] entails new primitive causal powers, reflected in laws which connect complex physical structures to the emergent features." (O'Connor & Wong, 2012) The problem here, strictly speaking, is that in accordance with the interpretation presented so far, there are no laws reflecting new causal powers and connecting complex physical structures to emergent features at the same time. On the one hand, L_{TO}-s, though connect complex physical structures to emergent features by assigning $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ and E to each other, do not reflect new causal powers since they express synchronic non-causal co-variation. On the

[29] Here I set aside questions concerning emergent-emergent causation and concentrate on emergent-physical causation. See Kim (1998, 2005) for an argument showing how emergent-emergent causation necessitates emergent-physical causation.

[30] First, due to limitations in space, I set aside the problem of whether the idea of effects temporally preceding their causes is viable or not. Second, note that cases where E and P_n are instantiated by a composite object and a constituent part of it, respectively, are the interesting cases of so-called reflexive downward causation. See §1.5.3 for more on the temporality of causes and effects.

other hand, L_{EC}-s, though reflect new causal powers by expressing causal connections between emergent and physical properties do not connect complex physical structures to emergent features since it is the emergent property *E* which is related to P_n by L_{EC} (*E*, P_n) not the complex structure $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$.

According to this interpretation, then, the instantiation of the structure of the physical properties $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ non-causally necessitates the instantiation of emergent property *E*, which in turn causes the instantiation of physical property P_n. That is, $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ plays an *indirect* role in the causation—first, the instantiation of $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ necessitates the instantiation of *E*, and second, the instantiation of *E* causes the instantiation of P_n. The picture this interpretation suggests is the following: due to L_{TO}, by instantiating a special structure of certain physical properties an object instantiates an emergent property which, due to L_{EC}, bestows new causal powers upon the object resulting in causing some novel physical effects.[31]

1.4.3 Understanding (E1)

So far, we have seen that there are two different sets of emergent laws: trans-ordinal laws and laws of emergent causation. Now let's consider what (E1) says about the status of these emergent laws. (E1) says that emergent properties are *basic* in the sense that they are not resultants of the physical realm. The ontological reading claims that non-resultant properties are metaphysically novel. They are not identical with any physical property (or set of physical properties). Moreover, they are metaphysically independent of the instantiation of physical properties: instantiating a non-resultant property is not equivalent with instantiating a certain set of physical properties. Therefore, (E1)—ontologically understood—claims that the instantiation of the appropriate set of physical properties in itself (given laws of physics) does not entail the instantiation of the emergent property. Emergent properties are basic, then,

[31] See §1.5.3 and §1.5.4 for more details about other possible interpretations of L_{TO}-s and L_{EC}-s.

in a relative sense that they are not *grounded*[32] in the physical realm (consisting of physical properties, plus physical laws).

However, though emergent properties are not grounded in the physical realm taken in itself, they are not totally ungrounded either, since they are grounded in the physical realm, *plus trans-ordinal laws*. This is what, according to our analysis, (E3) tells us: emergent properties—due to L_{TO}-s—depend on physical properties. It is due to the relevant L_{TO} that once $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ is instantiated E gets instantiated as well. This means, that emergent properties are non-resultants only if L_{TO}-s themselves are non-resultants (of physical laws and property distributions[33]). According to the ontological reading, then, it entails that trans-ordinal laws are fundamental laws in the same sense as those physical laws which are not grounded in other physical laws and property distributions. Were emergentism true of the actual world, in a possible world with the same fundamental physical property distribution and fundamental physical laws as in the actual world, emergent laws would still not necessarily hold. That is, according to emergentism, L_{TO}-s are not deducible (not even in principle) from, nor are they necessitated by the laws and property distributions of the physical realm.

Similarly, (E2)—in accordance with the ontological reading—claims that the new causal powers bestowed upon objects by emergent properties due to L_{EC}-s are novel in the sense that these causal powers are not grounded in the causal powers of the physical properties parts of these objects instantiate. Two possible worlds which are similar in every physical respect (having the same physical laws and same structures of physical properties) might be different with regard to the instantiation of P_n, i.e.

[32] Here I follow Jonathan Schaffer (2009) in taking ‘grounding’ as a primitive metaphysical relation which defines the notion of fundamentality—a property is fundamental if and only if it is ungrounded. Ungrounded properties are metaphysically primary; grounded properties are metaphysically dependent.

[33] And rules of composition. See Beckermann (2000) for the importance of rules of compositions.

with regard to whether laws of emergent causation hold or not. That is, LEC-s are not necessitated[34] by the laws and property distributions of the physical realm.[35]

To sum up, both LTO-s and LEC-s are independent of the laws and property distributions of the physical realm. In other words, LTO-s and LEC-s are *additional laws* in the sense that they neither are among the laws of physics nor are necessitated by them. As Samuel Alexander puts it concerning LTO-s, their existence is "to be noted [...] under the compulsion of brute empirical fact, or, as I should prefer to say in less harsh terms, to be accepted with the 'natural piety' of the investigator." (1920, ii, pp. 46-47).

1.4.4 Distinguishing physicalism and emergentism

Having all these at hand makes it possible to settle the issue of how to understand the 'nothing over and above' clause in the doctrine of physicalism. First of all, note that contrary to, for example, what Kim (2009) claims, ontological emergentism is a consistent view. True, on the face of it, (E1) and (E3) contradict to each other: whereas (E3) claims that there is a determination relation between physical and emergent properties, (E1) denies this. However, on the one hand, what (E3) claims is that there is a *nomological determination*—due to trans-ordinal laws—between emergents and the physical realm, whereas, on the other hand, what (E1) denies is that emergents are *grounded* in, i.e. are metaphysically determined by the physical realm in itself. That is, ontological emergence, as characterised by (E1)-(E3) is a coherent view.

Moreover, these characteristics readily distinguish emergentism from physicalism if we interpret the 'nothing over and above' clause as metaphysical determination. In

[34] The necessitation in question both here and in the case of LTO is of course metaphysical. The idea of laws being nomologically necessitated by other laws seems to yield the strange view that laws are connected by further laws (and thereby threatens with a regress).

[35] Nor are laws of emergent causation necessitated even by a base supplemented by trans-ordinal laws (and thus consisting of physical property distributions, rules of composition, physical laws, *plus trans-ordinal* laws). Cf. §1.4.2.

accordance with this interpretation, then, physicalism claims that all the facts are *metaphysically* necessitated by the set of physical facts and laws (cf. eg. Stoljar, 2005; Papineau, 2008). That is, what physicalists are committed to is that physical properties and physical laws together metaphysically necessitate all prima facie non-physical properties.

As we have seen, emergent properties are governed by two sets of laws: trans-ordinal laws (L_{TO}-s) and laws of emergent causation (L_{EC}-s). Both L_{TO}-s and L_{EC}-s are additional laws of nature in the sense that they are neither among the laws of physics, nor are necessitated by them. L_{TO}-s and L_{EC}-s are not grounded in physical laws and property distributions—they are additional in the *metaphysical* sense; two possible worlds identical in every physical respect might still be different with regard to how L_{TO}-s and L_{EC}-s look like, or even whether they are present. That is, contrary to what physicalism claims, physical properties and laws alone do not necessitate emergent properties. However, physical properties and laws *together with trans-ordinal laws* do necessitate emergent properties.

So physicalism is committed to and emergentism rejects metaphysical necessity in the following sense: prima facie non-physical properties are necessitated given physical laws and property distributions no matter what other laws of nature there might be. In what follows I shall argue that this is what ultimately tells physicalism and emergentism apart: whereas physicalists think that all properties supervene with metaphysical necessity on physical properties and laws emergentism holds that there are emergent properties which supervene only with nomological necessity on the very same base, since, in addition to physical properties and laws, emergent laws need to be fixed as well in order to get emergent properties.

1.5 A Problem with the Metaphysical/Nomological Distinction

Although differentiating between emergentism and physicalism on the grounds of the modal force of the necessitation relation they are committed to looks promising, it has to face with a recent objection put forward by Jessica Wilson (2005).

1.5.1 Jessica Wilson's objection

According to our analysis introduced above, if one copied all the physical laws and property-distributions of the actual world, but not the emergent laws, then (even if emergentism was true of the actual world) there would be no emergent properties and no novel causal powers in the copy world. Wilson (2005) argues in length that this analysis is mistaken. Her point is that without emergent laws being copied physical properties would not be instantiated. Wilson provides a twofold argument supporting this claim. On the one hand, she argues that the so-called Contingency view about laws—holding that "there are possible worlds where scientific properties of the type that actually exist are governed by very different laws than those actually governing such properties" (Wilson, 2005, p. 438)—is in tension with naturalism.[36] Since physicalists in general consider themselves as naturalists, this tension forces them to adopt the so-called Necessitarian view about laws over the Contingency view.[37] On the other hand, she argues that if one adopts the Necessitarian view—saying that the nature of properties depends on the laws of nature which govern them—then the conclusion above follows: without the emergent laws being present in a possible world the physical properties would not be properly instantiated. My aim here is to point out that Wilson's conclusion does not follow even if one accepts the Necessitarian view.

[36] What Wilson calls the Contingency view is a consequence of categoricalism, the view about the nature of properties endorsed in §1.4.1 and §1.4.2. Cf. footnote 26.

[37] Contrary to categoricalism, so-called *dispositional essentialism* claims that the roles properties play are *essential* to their nature. According to this view, laws of nature "spring from within the properties themselves" (Bird, 2007, p. 2). That is, dispositional essentialism is an anti-Humean position; it claims that laws of nature are metaphysically necessary. Wilson's Necessitarian view is essentialist about the causal roles properties play—cf. (S1) in Wilson's argument as reconstructed in §1.5.2. See §1.5.5 for more about dispositional essentialism.

According to the Necessitarian view: “any possible world where there exists a scientific property of a type that actually exists is a world where hold all the laws actually governing that property” (Wilson, 2005, p. 438). Consequently, if the nature of (at least some) physical properties depends on emergent laws, then in those possible worlds, which are a minimal physical duplicate of the actual world there hold all the emergent laws as well (given emergentism is true of the actual world). Thus the difference between physicalism and emergentism with respect to the modal force of the supervenience relation they endorse vanishes.

This argument might require some unpacking. Fist of all, Wilson subscribes to Gene Witmer’s view, according to which what is metaphysically possible reflects the nature of things under consideration (Witmer, 2001). Building upon this, Wilson argues that in those cases where the nature of entities under consideration is (at least partly) determined by the actual laws of nature the distinction between metaphysical and nomological possibility cannot be made. Wilson’s point is that the cases of physicalism and emergence are of this type, and thus it is pointless to argue that emergent properties are nomologically but not metaphysically determined by physical properties.

To drive this point home, Wilson needs to establish the claim that (some) physical properties depend on emergent laws. In doing so, she first introduces a line of thought supporting the Necessitarian view in general, and then she applies this scheme to the case of emergentism. Wilson reaches the general claim through the following steps. First, following (Shoemaker, 1980, 1998) and others[38] she accepts that the causal powers bestowed by so-called ‘broadly scientific properties’ essentially individuate these properties. Second, she adopts the view that laws of nature express (among other things) causal powers. From this, it follows that

[38] Eg.: Swoyer (1982), Elder (1994), Ellis (2001).

scientific properties are essentially individuated by their actual governing laws. (Cf. Wilson, 2005, p. 437)

The *Necessitarian view about laws* is what one gets after embracing all these steps. What this view tells us about emergentism is that if one copies all the physical laws and property-distributions of our world (supposing that emergentism is true of our world) but does not copy L_{TO}-s and L_{EC}-s, then the copy-world will not be indistinguishable from our world in every physical respect, since both L_{TO}-s and L_{EC}-s govern physical properties[39] which in turn, due to the lack of these emergent laws, will not be individuated properly. Strictly speaking, if emergentism is true then, according to the Necessitarian view, it is impossible to make an exact physical copy of the actual world without copying L_{TO}-s and L_{EC}-s as well. That is, in every possible world where the physical properties of the actual world are instantiated, emergent properties are instantiated as well—i.e. emergent properties supervene on physical properties with metaphysical necessity.

1.5.2 The core argument against Wilson's objection

For a detailed argument against this critical objection, let's first sum up how Wilson gets to her conclusion saying that emergent properties are metaphysically necessitated by physical properties. There are two crucial premises in the line of thought resulting in this conclusion. The first is the acknowledgement that there are emergent laws connecting emergent and physical properties. On the basis of our analysis of (E2) and (E3), we can say that this seems to be a well-established premise. The second premise is accepting the claim that certain physical properties are essentially individuated by these emergent laws. Wilson reaches this claim through the following steps:

(S1) Scientific properties are individuated by their causal powers.

[39] As it is captured by their formulations L_{TO} ($\langle P_1, P_2, \dots, P_k \mid \underline{s} \rangle, E$) and L_{EC} (E, P_n), where $P_1, P_2, \dots, P_k$ and P_n denote physical properties.

(S2) Laws of nature express causal powers.
(S3) Properties are individuated by their governing laws.

Here (S3), the claim in question, is the consequence of (S1) and (S2). If scientific properties are individuated by their causal powers (S1), and if these causal powers are expressed by laws of nature (S2), then these laws of nature expressing the causal powers of the scientific properties in question individuate (at least partly) scientific properties. Applying this general scheme to the special case of physical properties and emergent laws (S1)-(S3) read like this:

(S1*) Physical properties are individuated by their causal powers.
(S2*) Emergent laws express causal powers of physical properties.
(S3*) Physical properties are individuated by emergent laws.

That is, the statement that certain physical properties are essentially individuated by emergent laws relies on the truth of two claims. On the one hand, the truth of the general claim that all physical properties are individuated by their causal powers (S1*), and, on the other hand, the truth of the particular claim that there are emergent laws expressing causal powers of physical properties (S2*). In what follows, I presuppose the truth of the (S1)-(S1*) claim, and concentrate solely on the truth of (S2*).

For being able to evaluate the truth of (S2*), let's consider how it fits into our analysis of emergent laws. As we have seen, a trans-ordinal law $L_{TO}(\langle P_1, P_2,\ldots, P_k \mid \underline{s} \rangle, E)$ assigns a specific structure of certain physical properties and an emergent property to each other, expressing synchronic, non-causal co-variation: whenever the specific structure of certain physical properties $\langle P_1, P_2,\ldots, P_k \mid \underline{s} \rangle$ gets instantiated emergent property E also gets instantiated. It is not the case, that the instantiation of $\langle P_1, P_2,\ldots, P_k \mid \underline{s} \rangle$ causes the instantiation of emergent property E but rather that $\langle P_1, P_2,\ldots, P_k \mid \underline{s} \rangle$ non-causally necessitates E due to L_{TO}. That is, though L_{TO} act on physical properties $P_1, P_2, \ldots, P_k$ it does not express any of their causal powers. Trans-ordinal laws, then, do not satisfy (S2*).

Laws of emergent causation, on the other hand, do express causal powers. They express the novel causal powers of emergent properties and connect them to physical events. So a law of emergent causation L_{EC} (E, P_n) expresses that the instantiation of emergent property E causes the instantiation of physical property P_n. That is, L_{EC} (E, P_n) expresses forward looking causal powers of emergent property E, and backward looking causal powers of physical property P_n.[40] Note that L_{EC}-s expressing forward looking causal powers of emergent properties do not satisfy (S2*), since what (S2*) requires is an emergent law expressing causal powers of *physical properties*, and emergent properties are not physical properties.[41]

Where are we now? We have seen that for Wilson's argument to go through it is necessary for (S2*) to be true. (S2*) requires emergent laws to express causal powers of physical properties. There are two sets of emergent laws, L_{TO}-s and L_{EC}-s. L_{TO}-s do not express causal powers, only L_{EC}-s do. And the only physical properties whose causal powers are expressed by L_{EC}-s are P_n-s. In the general formulation of laws of emergent causation—L_{EC} (E, P_n)—P_n-s are placeholders for those novel physical effects which would not have occurred if emergent properties had not been instantiated. An important characteristic of emergentism is the very fact that emergent properties do make a difference in the sense that if a system instantiates emergent properties then its behaviour (the effects the system causes) will be different compared to a similar system not instantiating emergent properties. P_n-s are exactly the consequences of this difference-making. So when one chases the question

[40] See Shoemaker (2003, 2007) for the distinction between forward looking and backward looking causal features. Here I follow Jessica Wilson's terminology in characterising causal profiles with the expression of causal powers, instead of causal features. Accordingly, the backward looking causal power of P_n is to be understood as a part of the causal profile of P_n—namely its 'power' to be affected in a certain way.

[41] Note that I do not beg the question here. By the claim that emergent properties are not physical properties all I mean is that they are not properties of composing parts, i.e. they do not get into the supervenience base in question.

of how strong the dependency relation between an emergent property and a specific structure of certain physical properties (forming the supervenience base of the emergent property) is, what one definitely does *not* want to do is including P_n-s in the supervenience base. What this all means is that, though the causal powers (at least the backward looking ones) of P_n-s are expressed by L_{EC}-s, this has nothing to do with the question we are after here: whether the supervenience base of emergent properties determines emergents with nomological or metaphysical necessity—simply because P_n-s are not part of the relevant supervenience bases.[42]

The moral is that Wilson's argument doesn't go through since neither L_{TO}-s nor L_{EC}-s satisfy (S2*) in the relevant way: there are no emergent laws playing part in the individuation of those physical properties which form the supervenience base of the emergent properties in question. So there is nothing inconsistent in conceiving copying the supervenience base of a given emergent without copying the corresponding L_{TO}-s and L_{EC}-s. Referring to the difference in the modal force of the supervenience relation emergentism and physicalism are committed to, thus, remains to be a viable way of distinguishing between the views of emergentism and physicalism, and hence understanding physicalism as the doctrine that all facts are metaphysically necessitated by the physical facts (laws plus property distributions) seems to be the best way to understand it.

1.5.3 Refining the core argument—round 1: trans-ordinal laws as causal laws

In the previous section I have presented a core argument for the claim that Wilson's objection fails because neither trans-ordinal laws nor laws of emergent causation play a role in individuating those physical properties which form the supervenience

[42] It is possible to maintain the claim that P_n has no role here even if one wants to put this in terms of copying whole worlds instead of copying relevant supervenience bases only. Since the instantiation of P_n at t_1—in line with the diachronic understanding of causation—is caused by the instantiation of E at t_0 copying the whole world at t_0 does not copy P_n. In general, the causal relation between E and P_n might be synchronic or diachronic, and reflexive or irreflexive. The synchronic reflexive case would pose a problem even to the local supervenience based formulation (copying only the relevant base). However, I follow Kim (1999) in thinking that synchronic reflexive causation is problematic.

base of emergent properties. However, there are some loose ends which need to be tied up for my argument to succeed.

First of all, in answering Wilson's objection, I rely on the claim that trans-ordinal laws are not causal laws, but rather express synchronic non-causal necessitation. Were trans-ordinal laws causal laws, i.e. were the connections between $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ and *E* causal, trans-ordinal laws would satisfy Wilson's (S2*), since trans-ordinal laws would express causal powers of relevant physical properties. On the one hand, $P_1, P_2, \ldots, P_k$ form the supervenience base of emergent property *E*, thus their individuation is relevant here. On the other hand, if L_{TO} expresses causal powers of $P_1, P_2, \ldots, P_k$, then L_{TO} does play a role in individuating relevant physical properties.

Understanding emergentism as a causal phenomenon is certainly an option.[43] For example, most recently Timothy O'Connor (1994, 2000; 2005) has argued for a dynamical understanding of emergence where the relation between emergent properties and their base properties is diachronic and causal in nature. Nonetheless, it seems that such an approach is more the exception than the rule. E.g. McLaughlin (1992, 1997), van Cleve (1990), Crane (2001b), Kim (1999, 2006a, 2006b), and Papineau (2000, 2008) all agree that ontological emergence is best understood as a synchronic, non-causal determination relation.[44] It certainly seems right to say, that according to the received view—according to how it is most often interpreted—emergence is a non-causal phenomenon, i.e. trans-ordinal laws express synchronic non-causal determination. Nonetheless, as critiques might want to stress, the attractiveness of (or the amount of general sympathy towards) a causal understanding of emergentism is not an issue here. As long as a causal understanding is a viable and coherent option, it seems safe to say that Wilson's argument ultimately succeeds: the metaphysical/nomological distinction cannot serve as the basis of a *general* strategy for distinguishing emergentism and physicalism.

[43] See Mill (1843) for an early account where emergence is a causal phenomenon.

[44] For earlier versions of this position see C.D. Broad (1925) and Lloyd Morgan (1925).

However, the fact that the metaphysical/nomological distinction does not work for a causal understanding of emergentism is something I happily accept. Remember, the original question was whether the metaphysical/nomological distinction can differentiate between emergentism and physicalism. Wilson argues that it cannot since in cases where emergent laws govern physical base properties the metaphysical/nomological distinction simply vanishes. We have seen, that this conclusion only follows if emergent determination is interpreted as a causal connection, i.e. if L_{TO}-s are understood as causal laws. But under such an interpretation the original question does not even arise. The causal understanding of emergentism readily distinguishes itself from physicalism. Even if some interpret the determination relation within emergentism as a causal relation, no one has ever interpreted the determination relation within physicalism as such. Physicalism holds the view that prima facie non-physical properties are ultimately grounded in physical properties. Grounding is a synchronic relation. This clearly tells physicalism apart from any version of emergentism subscribing to a diachronic causal interpretation of the determination relation.

Moreover, interpreting emergent dependence as a *synchronic* causal connection won't help either. For physicalism also claims that the causal powers of prima facie non-physical properties are grounded in the causal powers of physical powers, and causal determination typically is not suitable for playing this role—the causal powers effects usually have are not grounded in the causal powers of their causes. That is, even if emergent determination might be understood as a synchronic causal connection, this version would straightforwardly distinguish itself from physicalism, since the determination relation physicalism is committed to cannot be interpreted in a similar way.

1.5.4 Refining the core argument—round 2: a different understanding of emergent causation

One might want to argue that the distinction between trans-ordinal laws and laws of emergent causation is unnecessary. Although it is true that LEC (E, P_n) connects emergent property E (as a cause) to physical property P_n (the effect), but since LTO ($\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$, E) provides a link between emergent property E and complex structure $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ by assigning them to each other, it is possible to substitute $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ for E in LEC (E, P_n). The resulting LEC* ($\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$, P_n) is a law which expresses the emergent novel causal power and connects it to a complex physical structure. One advantage of this approach might be that it allows one to commit oneself to some kind of an Occam's razor, and instead of positing two sets of additional laws (LTO-s and LEC-s) one can prefer positing only one set of additional laws (LEC*-s).

However, the strong claim behind merging LTO and LEC into LEC* is more than a simple application of Occam's razor. It yields an entirely new interpretation of emergence. According to this interpretation the formula LEC* ($\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$, P_n) expresses a causal connection between properties $P_1, P_2, \ldots, P_k$ and property P_n. It assigns a direct role to the physical properties $P_1, P_2, \ldots, P_k$ in causing P_n. That is, the strong claim behind the LEC* approach is that the physical properties $P_1, P_2, \ldots, P_k$ themselves bestow the causal power of instantiating the emergent effect P_n.[45]

Clearly, this approach poses a problem for my objection against Wilson's argument. For LEC*-s express novel causal powers of relevant physical properties ($P_1, P_2, \ldots, P_k$). This is exactly what is required by Wilson's crucial premise (S2*). So it seems, that under this interpretation of emergentism Wilson's argument has its full power

[45] Contemporary literature (e.g. O'Connor & Wong, 2012) often talks about 'emergent laws' connecting physical properties with emergent effects without distinguishing between trans-ordinal laws and laws of emergent causation—as though what they meant was something similar to the LEC* approach. However, as closer reflection reveals, they are not committed to the strong claim behind the LEC* interpretation as presented here. Rather they understand LEC* only as a shorthand for the LTO – LEC pair, assigning only an *indirect* role to $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ in the causation of P_n.

forcing us to abandon the metaphysical/nomological distinction as a useful one. However, I think that this is not the result Wilson really wants. Even if the metaphysical/nomological distinction does not apply to the LEC* approach, this leaves my argument entirely intact. Remember, the claim I am arguing for is that relying on the metaphysical/nomological distinction provides a general method for *distinguishing physicalism from emergentism*.[46] Now even if Wilson's argument goes through in this case, it doesn't really matter, simply because the LEC* approach is a non-starter here—since, as it turns out, it is a version of physicalism.

First of all, the LEC* approach eliminates emergent property *E*. That is, according to this view, there are no novel physically ungrounded emergent properties. Instead of *E*, physical properties $P_1, P_2, \ldots, P_k$ bestow the causal power of instantiating P_n upon the object in question. Moreover, there are no physically ungrounded causal powers either. The power of instantiating P_n might be novel, since none of the physical properties bestow it individually or in any combination other than $\underline{s}$. Nonetheless, the instantiation of P_n is grounded in the powers of the physical properties $P_1, P_2, \ldots, P_k$ instantiated by components of the object—namely (at least partly) in those powers which become manifested only if these properties get arranged into the specific structure $\underline{s}$.

What all this means is that the LEC* approach, by denying the existence of physically ungrounded higher level properties and physically ungrounded higher level causal powers, abandons the basic tenets of ontological emergence which made it incompatible with physicalism. To put it in another way, as some might see it, it is a virtue of the LEC* approach that it provides an understanding of emergentism

[46] In cases where telling physicalism and emergentism apart is not straightforward. Cf. §1.5.3.

compatible with physicalism.[47] The moral is that rather than being at odds with physicalism, the LEC* approach is a version of it. Thus there is no need for distinguishing them from each other. So the fact that we cannot differentiate between them on the grounds of the metaphysical/nomological distinction does not disqualify the strategy I am advocating.

1.5.5 Refining the core argument—round 3: property individuation and full-fledged dispositional essentialism

So far we have seen that Wilson's argument can't get off the ground since the requirements of premise (S2*) are not met—neither trans-ordinal laws nor laws of emergent causation express causal powers of relevant physical properties. However, the so-called Necessitarian view introduced in Section 3 is not as stringent as it could be. All it is committed to is that properties (at least partly) are individuated by their causal powers. This opens up the possibility of evading my objection by strengthening this commitment. If one claimed that properties are individuated not only by their causal powers but by *all* their *potentialities* then one would be able to block my argument.[48] Substituting the Necessitarian view with this latter view—which might be called 'full-fledged dispositional essentialism'—would entail (S1')-(S3') as follows:

[47] I do not have enough room here to discuss whether the LEC* approach provides a viable interpretation of emergentism. However, note that this view, as introduced above, is in fact an actual position held, for example, by Sydney Shoemaker (2007). Simplifying a bit, Shoemaker claims that $P_1, P_2, \ldots, P_k$ have two sets of causal powers: so-called *manifest* powers which they bestow upon the individual instantiating them regardless of the context of instantiation, and so-called *latent* powers which become manifested only if $P_1, P_2, \ldots, P_k$ get instantiated in specific structure $\underline{s}$ and remain latent when they get instantiated individually or in any other combination. The causal power of instantiating P_n then is grounded in the manifest and latent powers of $P_1, P_2, \ldots, P_k$. Shoemaker himself proposes this account as one which is compatible with physicalism. (Cf. Shoemaker, 2007, pp. 71-79)

[48] As it happens, what Wilson has in mind is a more stringent version of Necessitarianism (personal communication). She thinks that LEC (or as she prefers to think of it, the sudden occurrence of a new force field) is necessitated by the instantiation of $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$. As a consequence, in all possible worlds where $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$ is instantiated LEC is present. That is, Wilson's understanding of emergence is different from what is characterised by (E1)-(E3). Though Wilson is able to account for the novelty of emergent causal powers in some sense (in terms of the occurrence of a new force field) her view is unable to incorporate the claim that the causal powers bestowed by emergents are *genuinely novel* in the sense that they are not bestowed by any physical properties. This, then, threatens with Wilson's view being a non-starter as a version of ontological emergentism. See the argument of this section for more details.

(S1’) Physical properties are individuated by all their potentialities.

(S2’) Emergent laws express certain potentialities of physical properties.

(S3’) Physical properties are individuated by emergent laws.

My original strategy would clearly fail here, since the requirements of premise (S2’) —as opposed to those of (S2*)—are indeed met by trans-ordinal laws. For trans-ordinal laws do express certain potentialities of some relevant physical properties: namely that they give rise to emergent properties whenever they are instantiated in the right sort of way. According to full-fledged dispositional essentialism, $P_1, P_2, \ldots, P_k$ are individuated not only by their causal powers but also by all other potentialities they have, including the one that their instantiation in structure $\underline{s}$ brings about the instantiation of emergent property *E*. That is, the disposition that they give rise to *E* if instantiated in structure $\underline{s}$ plays a part in individuating $P_1, P_2, \ldots, P_k$, so in every possible world where they are present they have this disposition. In other words, there can be no possible worlds where $P_1, P_2, \ldots, P_k$ are instantiated but *E* is not, which means that instantiating *E* is metaphysically determined by instantiating $\langle P_1, P_2, \ldots, P_k \mid \underline{s} \rangle$. Thus the metaphysical/nomological distinction is of no use here in telling ontological emergentism and physicalism apart.

However, as others have already argued elsewhere (cf. Yates, 2009), this version of dispositional essentialism is at odds with any (E1)-(E3) incorporating interpretation of emergentism, especially with the claim that emergent properties bestow genuinely novel causal powers.

The problem is this. According to full-fledged dispositional essentialism, if an emergent property supervenes on a physical property then the fact that the physical property gives rise to the emergent property is essential to the physical property. Similarly: the genuinely novel causal power the emergent property bestows is essential to the emergent property. That is, what is essential to the physical property is that it gives rise to an emergent property bestowing that particular novel causal power (among other powers). Bestowing the novel causal power thus is, essential to the physical property. Had that particular causal power not been bestowed, the physical property would not have been individuated—for bestowing the causal power is necessary for individuating the emergent property, which in turn is necessary for individuating the physical property. By the definition of bestowal[49], then, it follows that the physical property bestows the novel causal power. This, however contradicts with how genuine novelty is understood in (E2). The causal power an emergent property bestows is supposed to be novel in the sense that it is not bestowed by the physical base of the emergent (neither is it grounded in the powers bestowed by the base).[50]

Where does this all leave us? First, we have seen that relying on the metaphysical/ nomological distinction seems to be a natural way of distinguishing physicalism and emergentism. Wilson's argument drew attention to the fact that the distinction cannot be made if relevant physical properties are individuated by emergent laws. I objected by showing that emergent laws do not play a part in individuating relevant physical properties. This objection fails though, if properties are individuated not only by their causal powers but by all their potentialities. Nevertheless, since those who subscribe to this latter view of property individuation cannot make sense of the emergentist claim that emergent properties bestow genuinely novel causal powers, the fact that the metaphysical/nomological distinction doesn't apply to this case leaves the

[49] There are different ways to formulate such a definition. The one favoured here claims that property *P* bestows causal power C upon object *o* if and only if, necessarily, for any object *o* which has causal power C, it has C *in virtue of* having *P*. The transitivity of the 'in virtue' relation, then, entails that the novel causal power of the emergent property gets bestowed by the physical property. Note the difference between this definition and a modal one defended, for example, by David Yates saying that a property bestows a power upon a particular if and only if, necessarily, if the particular instantiates the property then the particular has the power (cf. Yates, 2009, p. 119). The latter definition fails to correctly characterise cases where a particular instantiates two necessarily coextensive properties.

[50] This is not a problem for the categoricalist approach. For categoricalists can rely on laws of nature in accounting for *direct* power-bestowal. The categoricalist answer could then proceed in the following way. The emergent property directly bestows the novel causal power due to L_{EC}, whereas the physical property brings about the emergent property due to L_{TO}. The physical property bestows the causal power only in the sense of bringing about the emergent property. Neither L_{EC} nor L_{TO} connect the physical property to the causal power (and no other law does), thus on the categoricalist account, the physical property does not bestow *directly* the particular causal power in question. This law-based reply is not available for the full-fledged dispositional essentialist, neither are other strategies (cf. Yates, 2009, pp. 124-125).

original argument entirely intact. That is, relying on the difference in the modal force of the determination relation is still able to help distinguishing physicalism from any (E1)-(E3) incorporating interpretations of emergentism.

To sum it up, physicalism is committed to and emergentism rejects metaphysical necessity in the following sense: prima facie non-physical properties are necessitated by physical laws and property distributions no matter what other laws of nature there might be. That is, whereas physicalists think that all properties supervene with metaphysical necessity on physical properties and laws emergentism holds that there are emergent properties which supervene only with nomological necessity on the very same base, since, in addition to physical properties and laws, emergent laws need to be fixed as well in order to get emergent properties.

1.6 What is Physical?

Now that we have tracked how to understand the ‘nothing over and above’ clause in the definition of physicalism properly, our next task is to consider what the term ‘physical’ stands for. However, before turning our attention to this question, let’s first consider how metaphysical necessitation as featuring in the characterisation of physicalism might be best interpreted.

1.6.1 A note on how to understand metaphysical necessity: focus on the base set

If one considers the textbook definition of metaphysical and nomological necessity, then one might start to worry that contrary to all the efforts presented so far, one is forced to conclude that there is no difference in the modal force of the necessitation relation emergentism and physicalism embrace. For it seems that the textbook understanding of metaphysical and nomological necessity implies that both emergentism and physicalism embraces nomological necessity.

The problem is that taking the textbook definition of metaphysical necessity at face value seems to be in tension with how it is usually deployed in the physicalism literature. According to the general understanding of what metaphysical necessity is, something is metaphysically necessary if it is true in all possible worlds, *no matter what the laws of nature are in those worlds*. That is, something is metaphysically necessary in a possible world *w* if and only if it is the case in every possible world regardless of what laws hold in those worlds. Contrary to this, something is nomologically necessary, if it is true in all the possible worlds *with the same laws of nature as the actual world*. That is, something is nomologically necessary in a world *w* if and only if it is the case in every possible world with the same laws as *w*.[51]

In accordance with these textbook definitions, then, a physicalist endorsing metaphysical necessitation should claim that prima facie non-physical properties are determined by physical properties no matter what the laws of nature are in a given world. This, however, is clearly not the case: what physicalists are committed to is that physical properties and physical laws *together* necessitate all prima facie non-physical properties. The physicalist claim is that all God had to do in order to create everything there is was creating an initial distribution of physical properties plus setting up all the laws of physics.[52] Had a different set of physical laws been set up by God, the world would have turned out differently. Arguably, this is not what one has in mind when considering the textbook understanding of metaphysical necessity. Instead, it seems to be a standard case of nomological necessity: physical facts necessitate all the facts given laws of physics. That is, it seems that physicalism—

[51] On the one hand, the standard examples of metaphysical necessity are identities like 'Cicero is Tully' or 'water is H_2O' and cases of the determinable-determinate relation like the one holding between red and scarlet. If, for example, an object has the determinate colour scarlet then it has the determinable colour red, no matter what the laws of nature are. Similarly, in all possible worlds, regardless of the laws holding in those worlds, water is identical with H_2O. On the other hand, the standard examples of nomological necessity are cases like 'elephants can't fly', or 'water freezes at 32 degrees Fahrenheit'. Elephants cannot fly in those possible worlds where the laws of nature are similar to those of our world—it might well be the case that there are possible worlds where the gravitational coefficient is smaller than in our world and elephants can easily float around in mid-air. Similarly, were laws relevant for phase transitions different, water would possibly freeze at temperatures different from 32 degrees Fahrenheit.

[52] Cf. e.g. Crane (1991).

given the textbook definitions of metaphysical and nomological necessity—claims that prima facie non-physical properties are *nomologically* necessitated by physical properties.

This worry can easily be converted into an objection against the claim that it is the metaphysical/nomological distinction what tells emergentism and physicalism apart. Physicalism in fact, the objection goes, is a conjunction of two claims; first that prima facie non-physical properties supervene on physical properties with nomological necessity and second that there are only physical laws in the actual world. The emergentist agrees with the first claim, but denies the second, since according to emergentism, in addition to physical laws, there are also emergent laws in the actual world. That is, both physicalists and the emergentists agree that physical facts determine all the facts given laws of nature. They disagree, however, in what laws of nature there are. The consequence is that it is not the modal strength what distinguishes between physicalism and emergentism but the question of what counts as an emergent law and what counts as a physical law.

This problem, however, is only superficial. It draws attention to the fact that in the context of the physicalism-emergentism debate the term ‘metaphysical necessitation’ is utilised with a special focus on the supervenience base. According to what has been called the textbook definition, the supervenience base contains properties, and what is investigated is whether this base in itself is able to necessitate the supervening property. If it does, without the extra requirement of fixing laws of nature, then the necessity in question is metaphysical. However, there is a sense in which physicalists correctly refer to the necessity they are committed to as metaphysical—they claim that the supervening properties in question (the prima facie non-physical properties) are necessitated by a particular supervenience base without the requirement of fixing *additional* laws of nature. The difference lies in the fact that physicalists usually take the supervenience base to include, in addition to physical properties, the laws of physics as well. Thus, their commitment that fixing no extra laws of nature is required is exactly what rules emergentism out: no additional emergent laws are needed to be fixed for all the properties to be necessitated by the supervenience base in question (including physical properties and laws of physics). That is, even if, literally, the physicalists' use of metaphysical necessity collapses into the textbook use of nomological necessity (since it requires laws of nature, namely laws of physics to be fixed) it still seems to be correct to say that the supervenience base in question metaphysically necessitates what supervenes —correct in the sense that laws not in the supervenience base need not be fixed for this kind of necessitation to take place.

So what really happens when one evaluates supervenience claims like the one that physicalism is committed to is the following. First, one runs into prima facie non-physical properties and their governing laws. Then one considers whether prima facie non-physical properties supervene on physical properties and laws. This question ultimately comes down to whether the laws governing prima facie non-physical properties—i.e. LTO-s in the case of emergent properties—are determined by physical laws and property distributions. If they are, then they need not be fixed in addition to fixing the supervenience base of physical laws and properties, and physicalism turns out to be true. If, however, they are not, then fixing the supervenience base (physical properties and laws) is not sufficient on its own; the extra step of fixing additional laws of nature (the laws governing the prima facie non physical properties) is also necessary. In this latter case, physicalism turns out to be false.[53]

The moral is this: take a prima-facie non-physical phenomenon and consider whether it occurs in a possible world where the base-set (including physical laws and property distributions) is present without the necessary step of putting the governing

[53] Note that when considering the status of the governing laws of prima facie non-physical properties one can rely neither on the textbook understanding of nomological necessity—since as it is generally conceived there are no lawful connections between laws—nor on the textbook understanding of metaphysical necessity—since the necessitation does not hold regardless of laws of nature (what is at issue is exactly the status of certain laws relative to others).

laws of this prima facie non-physical phenomenon into the possible world as ‘extra’ or ‘additional’ laws—that is, consider whether the governing laws of the phenomenon in question are determined by the base set. Physicalists say that these governing laws are determined by physical laws and property distributions: once physical laws and property distributions are fixed, all the laws of nature are fixed. Emergentists, on the contrary, say that there are special properties whose governing laws are not determined by the base set.

1.6.2 Hempel’s dilemma and the *via negativa*

Now note that if it is true that emergent properties are nomologically necessitated—via emergent laws—by a supervenience base consisting of physical properties and physical laws, then it is also true that emergent properties are *metaphysically* necessitated by a supervenience base consisting of physical properties, physical laws, *plus emergent laws*. That is, if beyond physical properties and physical laws one also includes emergent laws into one’s supervenience base, then the above way of distinguishing physicalism from emergentism doesn’t work anymore. If physicalism is committed to the view that prima facie non-physical properties are metaphysically determined by a base set containing physical properties and physical laws, whereas emergentism is committed to the view that prima facie non-physical properties are metaphysically determined by a base set containing physical properties, physical laws, plus emergent laws, then it seems that the crucial difference between physicalism and emergentism is not the modal strength of the determination relation they endorse, but instead what they allow into the base set.[54]

[54] This line of thought stresses the point that emergent laws are just as fundamental as fundamental physical laws are. Both trans-ordinal laws and laws of emergent causation are independent of the physical realm—they are not necessitated metaphysically by the physical. The metaphysical independence of emergent laws might warrant why they should be included in the base set. As some could possibly argue, there is no difference between a fundamental physical law and an emergent law. Both are independent from any other set of laws and property distributions. To this effect, emergent laws are indistinguishable from fundamental physical laws. So if the base set should include all those properties and laws which are fundamental (that is, not determined by any sub-set of the base set) then it definitely should include emergent laws.

If, however, emergent laws were admitted into the base set, then the emergentist would agree with the physicalist that all prima facie non-physical properties are metaphysically determined by the base set. In this case, the only way to characterise the difference between physicalism and emergentism would be to define what to admit into the supervenience base. To put it in another way, what this line of thought asks is the following. Why do we refer, on the one hand, to certain fundamental laws as physical and include them in the supervenience base whereas, on the other hand, other similarly fundamental laws are dubbed emergent and additional and get excluded from the base?

The problem of how to describe what the term 'physical' amounts to in the definition of physicalism is called Hempel's Dilemma (cf. Hempel, 1969). Usually it is understood as a problem concerning properties—what types of properties should one rely on as constituents of one's supervenience base when formulating the doctrine of physicalism. However, the above line of thought shows that one should also consider laws, not just properties.

The most vivid exposition of Hempel's dilemma has probably been given by Crane and Mellor (1990) who argue that one cannot formulate what counts as 'physical' without making the term either obviously false or trivially true. In our present context, Hempel's dilemma is the problem of what one should include in the supervenience base of physicalism.[55] The first horn of the dilemma shows that if one defines the content of the supervenience base of physicalism on the grounds of current physics, then physicalism most certainly will turn out to be false, since it is

[55] Originally, Crane and Mellor concentrated on the causal closure of the physical, the crucial premise of the causal argument for physicalism. In that context Hempel's dilemma reads as follows. On the one hand, if 'physical' is explicated in terms of the entities current physics recognises, then the causal completeness of the physical is very unlikely to be true, since it is highly probable that future physics will identify new categories of physical causes. On the other hand, if 'physical' is the subject matter of a yet unknown future-physics, then either we are in no position to assess its closure, since we do not yet know what it is, or it makes the causal closure thesis trivial by presupposing the closure of the future-physics in question (cf. Crane & Mellor, 1990).

very likely that in the future—in the light of new scientific discoveries—physicist will revise the properties and laws currently posited by physics.[56]

The second horn of the dilemma warns us that relying on future completed physics won't help either. According to one formulation of this worry, a future *completed* physics will extend in such an extent that it will incorporate all properties. This future state of physics would then yield a quite literal sense of physicalism: there would be nothing over and above the physical, since everything would be included in the supervenience base (cf. Ney, 2008).[57] Admitting everything into the supervenience base naturally trivialises physicalism. It is more important from our present perspective what happens if one extends the supervenience base in a more economical fashion, i.e. if one admits only those properties and laws which are necessary for metaphysically determining all other phenomena. In this regard, only those properties and laws can get into the supervenience base, which are metaphysically independent of everything else already in the supervenience base and are necessary for metaphysically determining those phenomena which could not be so determined by the supervenience base prior to this extension. Nonetheless, this would still render physicalism trivially true: if physicalism is the view that all facts are metaphysically necessitated by a certain supervenience base, then defining what to admit into the supervenience base in a way that ensures that everything is metaphysically necessitated by the base makes the whole point circular.

[56] There are a few philosophers who nevertheless think that physicalism can rely on current physics. For example, Andrew Melnyk (1997) explicitly endorses this view. Moreover, it is possible to give a similar reading to Papineau (2000) as well. Papineau argues that there is a historic case for the causal closure in the sense that scientific discoveries of the past centuries has already shown us that the physical realm is causally closed. It seems to suggest that for Papineau current physics is causally closed. Since the causal closure of the physical is the crucial premise of the causal argument for physicalism, it follows that physicalism seems to be true even if one relies on current physics.

[57] There is another way of developing the second horn of the dilemma: if one tries to capture the idea of physicalism via relying on completed future physics, then it will become unclear what physicalism amounts to. After all, we just simply do not know what completed future physics will endorse.

The most popular way of overcoming Hempel's dilemma in contemporary literature is following a so-called *via negativa*[58] strategy, i.e. providing a 'complementary definition' by—instead of determining what *is* in the base set—determining what is *not* (cf. Spurrett & Papineau, 1999; Gillett & Witmer, 2001; Montero, 2001; Montero & Papineau, 2005; Worley, 2006). Typically, the *via negativa* strategy captures the physical as the non-mental. This move is motivated by an intuition stemming from the traditional antagonism between materialism and classical Cartesian dualism. Physicalism, which in this context is considered as the heir of materialism, should adhere to the thesis that mentality is not a fundamental (i.e. metaphysically independent) feature of the world. In this spirit, no matter to what extent future physics will extend the physical realm, it should not incorporate anything mental. That is, physicalism remains true as long as its supervenience base (metaphysically necessitating everything else) is free from mental properties and laws. This leaves us with the following definition: all facts are metaphysically necessitated by a supervenience base determined by ideal future physics, and containing nothing mental (cf. Wilson, 2006).

1.6.3 Conflicting intuitions

Mental phenomena, however, are not the only candidates for featuring in the *via negativa* strategy. David Papineau (2002), for example, argues in length that biological phenomena must also be excluded from the supervenience base of physicalism. Accordingly, for Papineau the physical is what is identifiable non-mentally *and* non-biologically, i.e. as inanimate (cf. Papineau, 2002, pp. 40-43). Beyond mental and biological phenomena other candidates for being excluded from the supervenience base in the definition of physicalism are chemical, social, ethical, aesthetic, etc. phenomena. Excluding only the mental is worrisome since fundamental chemical, biological, social, etc. phenomena are just as demolishing with regard to physicalism as fundamental mentality is: if fundamental chemical, biological, social, etc. properties and laws were needed for accounting for all the

[58] This term has most probably been coined by Gillett and Witmer (2001).

facts (in a metaphysical sense), then that would just as much falsify physicalism, as if fundamental mentality was needed. In other words, excluding only the mental from the supervenience base is too arbitrary, and does not cover the original connotations of physicalism.

Such a worry might be motivated by another intuition capturing the tension we could already observe in the previous sections: physicalism is incompatible with ontological emergentism. Applying the *via negativa* strategy in this context amounts to determining the supervenience base of physicalism as ‘non-emergent’: if one is persuaded by the intuition that physicalism is incompatible with ontological emergence, then one cannot admit anything into the supervenience base of physicalism, which is claimed to be ontologically emergent.

The worry that excluding only the mental from the supervenience base is an arbitrary move draws attention to the fact that following the ‘physicalism is not Cartesian dualism’ intuition is in tension with the ‘physicalism is not ontological emergentism’ intuition, since holding that physicalism and ontological emergentism are incompatible is a stronger commitment than holding that physicalism and Cartesian dualism are incompatible. Imagine, for example, that certain chemical phenomenon turned out to be ontologically emergent. In this case, for those who are guided by the ‘physicalism is not ontological emergentism’ intuition, physicalism would be falsified, since they would feel inclined to exclude the chemical from the supervenience base, which in turn would not metaphysically necessitate all the facts. Contrary to this, those who prefer only the ‘no fundamental mentality’ clause, and thereby happily admit fundamental chemical properties or laws into the supervenience base would see this only as a necessary extension of the base (in order to ensure that it metaphysically necessitates all the facts), and thus would still feel that physicalism is safe and sound.

Interestingly, Jessica Wilson, though shares the intuition that ontological emergentism is incompatible with physicalism (Wilson, 2005), nevertheless thinks that relying on the ‘no fundamental mentality’ clause suffices (Wilson, 2006). Wilson argues that one does not need to be bothered with candidates other than the mental since either excluding the mental readily excludes them as well, or they have already been proved to be ontologically dependent on the physical. On the one hand, mentality is a precondition for the existence of social, ethical, aesthetic (etc.) phenomena. It is very plausible, Wilson claims, that social processes, moral agency, aesthetic response are all, at least to some degree, constituted by mentality (cf. Wilson, 2006, p. 76). On the other hand, as the advances of the last century testify, chemical and biological phenomena can be metaphysically necessitated by physical entities. As Wilson puts it: “chemical and biological features of reality can, in actual fact, be ontologically accounted for in terms of configurations of relatively fundamental entities that are not themselves chemical or biological” (Wilson, 2006, p. 75).

That is, Wilson argues that even if we are committed to the intuition contrasting physicalism with ontological emergentism we have no reason to worry about other than the mental. Excluding the mental from the supervenience base of physicalism will automatically exclude the social, ethical and aesthetic, whereas, in line with current science, it seems to be a safe bet to say that there are no ontologically emergent chemical and biological phenomena.

Note, however, that something must have gone astray here, with regard to the latter case. The issue at hand is not what current science seems to suggest, but rather what we need to pre-cautiously exclude from the supervenience base in order to preserve the intuitive message of physicalism no matter what future science will discover. The very idea behind the *via negaitva* strategy is that determining which natural phenomena are such that their fundamentality would intuitively falsify physicalism, and defining the physical accordingly, gives us a tool to block the worry that not yet known future physics could unacceptably extend the range of the physical, and hence render the doctrine of physicalism trivially true.

It very well might be the case (even if at the moment it seems improbable), that at some point future science will discover certain biological or chemical processes and claim that they are fundamental—that they cannot be ontologically accounted for in terms of, and are metaphysically independent from the non-biological or the non-chemical. This possible future development of science would intuitively falsify physicalism. Excluding fundamental chemical and biological properties and laws from the supervenience base encodes this intuition into the definition of physicalism.

Note that this worry is not entirely groundless even from the perspective of present day science. For example, the question whether there are emergent biological phenomena is still debated. Even if the vast majority of the physicalism literature agrees that emergence within biology (if any) is only epistemological, some still find it possible that there are genuinely (i.e. ontologically) emergent biological phenomena (cf. eg. Boogerd, et al., 2005; Kauffman & Clayton, 2006). Similarly, some still think that the arguments which are supposed to show that the chemical can be ontologically accounted for in terms of the physical are inconclusive (cf. eg. Scerri, 2007). Moreover, some argue that there is ontological emergence even within physics. E.g. Silberstein and McGeever (1999), Laughlin and Pines (2000), Batterman (2002), and Papineau (2008) all hold the view that there are (or at least might be) ontologically emergent macro-physical phenomena.[59]

[59] The latter is an interesting issue. Ontologically emergent macro-physical phenomena would clearly falsify *microphysicalism*—the view that all the facts are metaphysically necessitated by the properties and laws of the mereologically most fundamental entities (cf. Pettit, 1993, 1994; Papineau, 2008; Tye, 2009). However, it would be entirely compatible with versions of physicalism which are not committed to the intuition that the domain of the 'physical' in the definition of physicalism should be limited to the properties and laws of the mereologically most fundamental entities. See Papineau (2008) for a detailed discussion.

The moral is that Wilson is too quick when she takes it for granted that emergentism about the chemical or the biological has already been proved wrong. Moreover, and more importantly, she seems to miss the whole point behind the *via negativa* strategy.

1.6.4 The ultimate *via negativa*: physicalism relativised

Recall that ontological emergentism formulated as a general thesis tells us that there are certain special (emergent) properties which are necessitated nomologically but not metaphysically by the physical. This formulation, however, suffers from the very same problem the general formulation of physicalism struggles with: it needs to determine what the 'physical' is.

The *via negaitva* strategy provides an easy answer here: a particular property is ontologically emergent if it is necessitated nomologically but not metaphysically by a supervenience base including neither the particular property in question nor its governing laws. In fact, proponents of emergentism are always emergentists *about a particular phenomenon*.[60] One might be an emergentist about certain biological phenomena, certain chemical phenomena, certain physical phenomena, etc. Note that these are quite different varieties of emergentism. Emergentism about the chemical, for example, claims that there are certain chemical properties which emerge out of a base set containing nothing chemical. Emergentism about the biological, on the other hand, claims that certain biological properties emerge out of a base set which might contain something chemical but definitely nothing biological. Even if Wilson is probably right that excluding the mental from the supervenience base will exclude everything social, ethical, aesthetic, the opposite is certainly not true: a supervenience base including the mental must not automatically include everything social, ethical or aesthetic. Therefore emergentism about, say, the social could also be a coherent view claiming that there are social phenomena emerging out of the non-social (which might very well include individual mental properties and laws).

[60] Cf. Broad and Scerri, who are emergentists about certain chemical processes (Broad, 1925; Scerri, 2007), or Laughlin and Batterman, who are emergentists about certain physical phenomena (Laughlin & Pines, 2000; Batterman, 2002), etc.

These different versions of emergentism tell us something *about different target domains relative to different base domains*: the properties and the laws of the particular target domain under consideration are emergent relative to the base domain, where the only restriction regarding the base is that it should not include the properties and laws of the actual target domain. That is, the right way of thinking about emergentism is not as *emergentism per se*, i.e. as a single, general doctrine, but rather as *emergentism about something*. Emergentism is not an absolute notion but a relative one—relative to the target and relative to the base.

Note that all that has been said so far is in line with how the whole issue of emergentism, and for that matter, how the whole issue of physicalism arises. First, one runs into a phenomenon, which presents itself as a special phenomenon with some characteristic properties. On the basis of these characteristic properties one categorises it as, for example, a chemical, biological, or mental (etc.) phenomenon. This categorisation also means that the particular phenomenon will present itself as a *prima facie* non-physical phenomenon, since its characteristic properties intuitively distinguish it from standard physical phenomena.[61]

The question to be answered then is whether these properties and their governing laws are metaphysically necessitated by an appropriately chosen base set. Note that including these properties or their governing laws into the base set just begs this question. Consider how uninteresting it is to ask if mentality is metaphysically necessitated by a supervenience base containing mental properties. Of course it is. The interesting question is whether mentality is metaphysically necessitated by a supervenience base *not* containing anything mental. If the answer is no then dualism about the mental will follow. If the answer is no, but there are psychophysical laws of the trans-ordinal kind (i.e. if the mental is nomologically necessitated), then ontological emergentism about the mind will follow. Finally, if the answer is yes,

[61] Such characteristic properties and the related non-physical categories might be, for example, intentionality for the mental, or agenthood for the biological.

then what one gets might be called *physicalism about the mental*. That is, the moral above—namely that emergentism (and also dualism and most interestingly from our present perspective, even physicalism) is a relative notion is exactly what is dictated by the logic of how the problem at issue arises.

Note how the original moral has been generalised. It is not just ontological emergentism which is naturally reads as a relative notion. As we have seen it in the previous paragraph, dualism and even physicalism can be given a similar, relative reading, defining it as a particular claim *about a certain target phenomenon* and *relative to a base not containing anything specifically related to that very target*.

Relative physicalism about a certain phenomenon, thus, is the thesis that a particular target phenomenon is metaphysically necessitated by a supervenience base from which all characteristic properties and laws of the target phenomenon are excluded.[62]

That is, from this perspective it seems that the ultimate intuition motivating physicalism is this: if a phenomenon is characterised by some special, *prima facie* non-physical features then one does not need to posit these characteristic features as fundamental for being able to ontologically account for the target phenomenon.[63]

Relative physicalism, in a certain sense, is significantly different from the general doctrine of physicalism. Nonetheless, it gets the spirit of physicalism right: the metaphysical fabric of nature is tight. Phenomena which are identified via certain

[62] And also the proto-properties of the phenomenon—cf. e.g. naturalistic dualism (Chalmers, 1996), or panpsychism (e.g. T. Nagel, 1979). Of course, the base set might be further restricted. One might be interested in the question whether a particular phenomenon can be ontologically accounted for in terms of the properties and laws of the mereologically most fundamental entities, which then would yield *microphysicalism about that particular phenomenon*.

[63] This claim captures why it seems intuitively attractive to exclude anything mental, chemical, biological, social, etc. from the supervenience base of physicalism considered as a general doctrine—the related characteristic features (e.g. intentionality, chemical behaviour, biological agency, social dynamics, etc.) are those which *prima facie* appear to be non-physical, i.e. are identified as features different from those physical entities typically possess. Cf. Crane (2001a) who tries to overcome the second horn of Hempel's dilemma by claiming that all that physicalism needs is an ontological commitment to that *kind* of entities current physics posits. The *prima facie* non-physical features in question here are typically of a very different kind than those posited by current physics.

characteristic features and thus present themselves as *prima facie* non-physical are never fundamental. The metaphysically necessary building blocks all belong to the same domain—there are no other domains floating around freely, without being metaphysically anchored. That is, all seemingly independent domains of nature are in fact metaphysically connected to each other.

In accordance with all this, from now on the version of physicalism, which is going to be in the centre of my interest is *physicalism about consciousness and relative to a base containing nothing conscious*. This supervenience base, beyond physical properties and laws, also contains chemical and biological phenomena. That is, from now on, I shall use the term 'physical' as a cover-term for the physical, chemical and biological. In fact, certain characteristically biological processes play a crucial part in the account that shall be introduced in Chapter 4.

Chapter 2:
Physicalism about Consciousness and the Epistemic Gap

2.1 Conscious Experience

Imagine that you wake up, open your eyes, and gradually become aware of your surroundings. You see the face of your loved one lying on the pillow right next to you. You hear her soft breathing, and smell the scent of her hair. These feelings are part of how you experience her. In fact, the way the shape of her face, the sound of her breathing, and the smell of her hair appear to you, are distinctive characteristics of your actual conscious experience.

The problem of consciousness—which is the immediate context of this doctoral thesis—is whether the phenomenon that can be grasped via such characteristic features fits into the physical world. *Physicalism about consciousness* is a thesis claiming that it does: conscious experience is metaphysically necessitated by a supervenience base excluding everything having these characteristic features.[1]

Physicalism about consciousness is motivated by a particular version of the causal argument for physicalism introduced in §1.2. It is one of the most fundamental observations we have that our conscious experiences are tightly embedded in the causal network of the physical world. Physical causes do affect our conscious experiences, which in turn are also able to affect the physical world. A hornet's sting (a physical cause) results in sharp pain (a conscious experience) which then affects our bodily movements—we pull our hand away immediately (a physical effect). As David Papineau summarises it: "conscious mental occurrences have physical effects" (Papineau, 2002, p. 17). This observation, once plugged into the causal argument as its first premise, leads (via the 'causal closure of the physical' and 'the

[1] Or their proto-variants. Cf. Footnote 62 in §1.6.3.

no-overdetermination' theses) to the conclusion that conscious mental occurrences must be identical with physical states.

However, what makes the case of consciousness more like a mystery is that at the same time we undoubtably have a quite strong intuition pulling us towards the opposite direction. Shapes as we see them, sounds as we hear them, scents as we smell them—i.e. the distinctive characteristics of our experiences, which seem to define the very nature of consciousness—are nothing like physical states.[2]

2.1.1 The phenomenal character of conscious experience

The world we live in is the world of our conscious experiences. Everything we perceptually know about the environment around us is given as a conscious experience. When I wake up and see *that* face, hear *that* breathing, smell *that* scent, and recognise where I am and who is lying next to me, there is *something it is like for me* to be there, and see, hear and smell all those things. I live in a vivid world of feelings.

In his seminal paper, Thomas Nagel (1974) introduced the notion of what-it-is-like-ness, and argued that what having a conscious experience for a creature amounts to is there being something it is like to be that creature. In other words, a creature is conscious (have a conscious experience) if there is something it is like for that being (to undergo the experience in question). The world appears in a certain way to those creatures which are capable of having conscious experiences—they have a rich inner phenomenal life. This inner life is what is often called the phenomenal or the

[2] Cf. the characteristic features of prima facie non-physical phenomena as discussed in Footnote 61 and Footnote 63 in §1.6.3.

qualitative character of conscious experience.[3] Phenomenal properties, or qualia, are the introspectively accessible features of the inner life—the way things appear to those having the experiences.[4]

Note that it is not just the case that phenomenal properties are introspectively accessible, but they are accessible *only* introspectively.[5] The phenomenal character of an experience is exactly what it is like *for the creature actually experiencing it* to have that particular experience. No other creatures have access to that very token phenomenal quality. That is, the phenomenal character of an experience is a *subjective* character. Even if I react by saying 'ouch' and seem to know what it could be like when my friend is telling me about her being stung by a hornet, this is just due to the fact that I suppose that my friend's pain is similar to my pain. I do not have access to her pain; by saying 'ouch' I do not respond to her experience. I am simply empathising with her—thinking of my own pain and projecting it on her.[6] It is especially obvious if we consider creatures whose perceptual system is different from ours. Nagel's original examples were bats (T. Nagel, 1974). Bats use sonars to navigate, which is unlike any perceptual modality we humans have access to. This fact in itself prevents us from being able to rely on presumed similarities between our

[3] Note the difference between creature- and state-consciousness. A creature (i.e. a whole organism) is conscious if there is something it is like to be that creature, i.e. if there is a specific way the world appears from the point of view of the creature. A particular mental state or process is conscious if there is something it is like to be in that state, i.e. if the state has qualitative or phenomenal properties. Phenomenal properties are sometimes thought of as referring to the overall structure of experience. In this sense, the qualitative character of experience, the 'raw feeling', is a constituent of the phenomenal character—together with the spatial, temporal, and conceptual organisation of experiences (cf. Van Gulick, 2011).

[4] Since Nagel published his paper, it has become quite a custom to appeal to the what-it-is-like-ness of an experience in order to grasp the fundamental characteristic of phenomenal consciousness (cf. e.g. Chalmers, 1996; Crane, 2001a; Papineau, 2002). See, however, Paul Snowdon (2010) for an argument claiming that Nagel's slogan is, in fact, trivial and absolutely uninformative.

[5] I am not committed to any particular account of introspection here. See Schwitzgebel (2010) for more on different approaches to introspection.

[6] To be more accurate, it is *non-inferential access* to the phenomenal character of an experience that is restricted to the actual subject of the experience in question. It is possible to have inferential access to one's pain, for example, via observing one's pain behaviour. The issue here, however, is not accessibility in this broad (inferential) sense, but rather *ownership*. A particular experience necessarily belongs to the actual subject—it is owned by the subject. Each particular experience must have an owner, and it can be owned only by that very subject actually experiencing it. Cf. e.g. Tye (1995, 2007).

experiences and theirs. Therefore, we have no idea what it is like to be a bat. In other words, the phenomenal character of conscious experience is inherently tied to a particular point of view—the point of view of the subject of the experience. Phenomenal consciousness, as it might be put, is the first-person aspect of the mind (cf. Chalmers, 1996, p. 16).[7]

2.1.2 The hard problem

Let's consider again the case of me waking up and finding myself in the middle of a phenomenal world consisting of shapes, sounds and smells. These experiences help me determining where I am and recognising who is next to me. However, I can imagine a fancy automaton which would just as well be capable of detecting the light rays reflected from the face, the air-waves produced by the breathing, and the molecules emitted by the hair. This automaton, by analysing the signals detected, could even determine its location and recognise the person next to it. However, no inner life it would have—there would be nothing it would be like for the automaton to detect all those signals. This automaton, though could process information just as well as we do, would lack the inner life so characteristic of ourselves. It seems that information processing and phenomenal consciousness are two distinct phenomena.[8]

Several authors distinguish between two different kinds of consciousness along these lines. David Chalmers, for example, contrasts the phenomenal concept of the mind with the psychological concept of the mind (Chalmers, 1996), whereas Ned Block argues for a distinction between phenomenal and access consciousness (Block, 1995). Both of these dichotomies try to capture the distinction between a first-person, qualitative, introspectively accessible, and an objective, causal, scientifically

[7] Note that being aware of the phenomenal qualities of a conscious experience amounts to being aware of the qualities of the object the experience represents. That is, these qualities do no appear to be the qualities of the experience itself, but rather qualities of the external object (or bodily process). The shape of the face, the sound of the breathing, the smell of the hair one is aware of after waking up all appear to be qualities of the face, the breathing and the hair respectively, and not of the corresponding experiences themselves. That is, experiences are *transparent*, or *diaphanous* (cf. e.g. Tye, 2000, 2007).

[8] See Chapter 4 and Chapter 5, and especially §4.3.3, §5.3.2 and §5.3.3 for my own view about the relation between information processing and phenomenal consciousness.

expressible aspect of consciousness. On the one hand, Chalmers' phenomenal concept of the mind and Block's phenomenal consciousness both refer to the subjective phenomenal character of conscious experience as introduced in §2.1.1. On the other hand, the psychological concept of the mind and access consciousness are terms trying to grasp the third-person, information processing aspect.

Chalmers clarifies what he means by the psychological concept in this way:

> "This is the concept of mind as the causal or explanatory basis of behavior. A sate is mental in this sense if it plays the right sort of causal role in the production of behavior, or at least an appropriate role in the explanation of behavior. According to the psychological concept, it matters little whether a mental state has a conscious quality or not. What matters is the role it plays in a cognitive economy." (Chalmers, 1996, p. 11)

Block gives the following three conditions of access consciousness, which he takes to be together sufficient, but not all (especially the third one) necessary:

> "A state is access-conscious if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous [...], that is, poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech." (Block, 1995, p. 231)

'Access consciousness' and 'the psychological concept of the mind' are functional notions—what makes a state access conscious or mental in the psychological sense, is what the sate does (what functional role it plays) in the system it is embedded in. Contrary to this, the phenomenal character of conscious experience seems not to be a functional notion.[9] To see this, consider again our model case: seeing the face, hearing the breathing and smelling the scent of the hair of the loved one lying next to us. Seeing, hearing and smelling are all functional notions (standing for processes consisting of how external stimuli get detected, how categorisation and recognition happens, etc.). That is, in order to explain these capabilities what one needs to do is

[9] Though see, for example, Dennett (1991) who argues in length that consciousness is a purely functional notion.

specifying certain mechanisms which are able to perform the functions in question. Once this is done, an explanation of how the information processing goes during these activities is given. But why this information processing is accompanied by a phenomenal feel—i.e. why it is like something to see, hear and smell—seems to be a further question. All these activities could go on 'in the dark' without the presence of any inner life, just like in the case of the automaton above. That is, explaining all the functions leaves out the phenomenal character of consciousness.[10]

This idea is already present in Nagel's original text. As he puts it:

> "[The what-it-is-like-ness of an experience] is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing. It is not analyzable in terms of the causal role of experiences in relation to typical human behavior—for similar reasons." (T. Nagel, 1974, pp. 436-437)

Now the problem is that the way science typically progresses is by providing functional explanations. The way science helps us understand that, say, water is H_2O consists in determining the functional role water plays, and then showing how H_2O is able to fill the very same functional role.[11] If, however, the realm of the phenomenal character of conscious experience is left intact by functional explanations, then standard scientific methodology seems to be inapt to help us understand how phenomenal consciousness is connected to the physical world. This is why Chalmers calls the problem of experience the *hard problem* of consciousness (Chalmers, 1995). As opposed to the so-called easy problems (e.g. explaining discrimination, reportability, etc.), in the case of the hard problem we do not even know how to start tackling the problem. The so-called easy problems might, of course, be relatively

[10] For more on functional explanations leaving out the phenomenal aspect of consciousness see §2.2.2. See also Chapter 6 and Chapter 7.

[11] To be more precise, science provides many different kinds of explanation, e.g. causal, teleological, reductive, etc. However, it is functional explanation which is in the centre of attention in the debate over physicalism about consciousness. See Chapter 6 for a detailed discussion of this issue. Note, that science also provides structural descriptions. The significance of this is introduced in §4.1.2.

hard, but at least we have a grip on the subject matter: we know how to approach the question with our standard scientific methodology. Contrary to this, however, we are clueless with regard to the hard problem.

2.2 Epistemic Arguments against Physicalism about Consciousness

Explaining why certain physical processes are accompanied by phenomenal qualities could very well be a hard problem—this in itself, however, would not be fatal for physicalism about consciousness. Nonetheless, some think that accounting for phenomenal consciousness is not just a hard but an intractable problem for physicalism. They argue that there are sound arguments showing that physicalism about consciousness is in fact false.

2.2.1 The general strategy

Such arguments usually proceed as follows. First, they establish an epistemic gap between the realm of our conscious experiences and the realm of physical processes. Second, from the presence the epistemic gap, they infer the existence of an ontological gap between physical properties and phenomenal properties. Third, they point out that if there is an ontological gap between physical properties and phenomenal properties characteristic of conscious experience, then physicalism about consciousness is false.

The first step, establishing an epistemic gap, amounts to showing that there is no *epistemic entailment* from physical/scientific knowledge to phenomenal knowledge. In other words, there is no epistemic entailment[12] from physical truths to phenomenal truths. Signs of this so-called *epistemic gap* has already appeared in the previous

[12] Cf.: "the most basic sort of epistemic entailment is a priori entailment, or *implication*. On this notion, *P* implies *Q* when the conditional 'If *P* then *Q*' is a priori—that is, when a subject can know that if *P* is the case, then *Q* is the case with justification independent of experience" (Chalmers, 2010a, p. 109). Chalmers argues that all the different versions of the epistemic argument against physicalism (cf. §2.2.2-§2.2.4) can be interpreted as making a case against the claim that physical truths imply phenomenal truths.

section. We have seen that standard scientific methodology is inapt to help us understand why there is conscious experience accompanying physical processes. No matter how detailed our knowledge is in terms of physical properties, we cannot infer from this phenomenal truths like what it is like to undergo a certain conscious experience (cf. §2.2.2). Similarly, as we have also seen, even if a creature is capable of all the fine details of information processing we perform, we still find it conceivable[13] that the creature is only an automaton, i.e. lacks an inner phenomenal life (cf. §2.2.3).

The second step of the general strategy consists of concluding on the existence of an ontological gap on the basis of the presence of an epistemic gap. This is the crucial premise of the epistemic arguments—in fact, this is the premise doing all the work in the process of reaching the desired conclusion that physicalism is false. Without it, pointing out that there is an epistemic gap between physical and phenomenal knowledge would remain a claim purely about knowledge itself, without any metaphysical conclusion. Claiming that there is an ontological gap between physical properties and phenomenal properties amounts to arguing for the failure of *ontological entailment* between physical and phenomenal facts. Physical facts fail to ontologically entail phenomenal facts if phenomenal properties are not necessitated metaphysically[14] by physical properties, e.g. if it is possible for the physical facts to hold without the phenomenal facts holding (cf. Chalmers, 2003, 2010a).[15]

Physicalism about consciousness is the thesis that conscious experience is metaphysically necessitated by a physical supervenience base, i.e. a base excluding

[13] The kind of conceivability in question here is ideal negative conceivability: a statement *S* is conceivable in this sense if a reasoner cannot rule out *S* on the basis of ideal rational reflection. That is, if *S* is conceivable, then *S* is epistemically possible. It is the crucial second step of the epistemic arguments against physicalism which try to establish a further link between conceivability and *metaphysical possibility* (cf. Chalmers, 2002).

[14] Cf. Chapter 1, especially §1.5.

[15] The move from an epistemic gap to an ontological gap is motivated by the so-called *Entailment Thesis*. See §2.3 for a detailed discussion. See Chalmers (2002, 2004, 2010b) for an extensive argument (so-called two dimensional argument) supporting the link between conceivability and metaphysical possibility.

everything having the characteristic features of conscious experience (cf. §1.6.4). Therefore, if there was an ontological gap between the physical and the phenomenal, then physicalism about consciousness would clearly be false: the ontological gap between the physical and the phenomenal would prevent the physical supervenience base to metaphysically necessitate conscious experience. This concludes the epistemic argument against physicalism about consciousness.

That is, the common structure shared by all the arguments which start from the observation that there is an epistemic gap between physical truths and phenomenal truths, and reach the conclusion that physicalism about consciousness is false can be summarised as follows:

Epistemic Argument against Physicalism about Consciousness

(P1) There is an epistemic gap between physical and phenomenal truths.

(P2) If there is an epistemic gap between physical and phenomenal truths, then there is an ontological gap between physical and phenomenal properties.

(P3) If there is an ontological gap between physical and phenomenal properties, then physicalism about consciousness is false.

(C) Physicalism about consciousness is false. (Cf. Chalmers, 2010b, p. 110)

2.2.2 Levine’s explanatory gap

Now that we have seen the general structure of the epistemic arguments against physicalism about consciousness, it is time to discuss the individual arguments in detail. First, consider the so-called Explanatory Gap Argument.

This argument starts with an analysis of one of the observations already introduced in §2.1.2, namely that explaining all the functions seems to leave out conscious experience. Joseph Levine (1983) compares identity statements typically populating scientific explanations[16] like ‘water is H_2O’, or ‘heat is mean molecular kinetic energy’ with identity statements concerning consciousness like ‘pain is C-fibres firing’. He argues that whereas scientific identity statements are *fully explanatory*, psycho-physical identities concerning consciousness leave something crucial unexplained. The identity statement ‘heat is mean molecular kinetic energy’ is fully explanatory in the sense that:

> “our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role [...] exhaust our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand.” (Levine, 1983, p. 357)

Of course, as Levine himself acknowledges it too, the claim that ‘pain is C-fibres firing’ can be explanatory in a very similar sense. One might feel tempted to argue from one’s subjective point of view, that one winces and pulls one’s hand away *because* of the pain one feels when one is stung by a hornet. That is, ‘pain’ can be associated with a causal role (e.g. producing pain behaviour upon receiving painful stimuli). Then, by knowing how C-fibres firings fill this role, i.e. by knowing the mechanism how the hornet’s sting activates C-fibres, and how C-fibres stimulation causes wincing and other pain behaviour, one becomes able to motivate, to make intelligible the ‘pain is C-fibres firing’ claim. In this sense, the statement ‘pain is C-fibres firing’ is explanatory.

However, it is not *fully* explanatory. The concept of pain is only partly of the causal role of pain—it is also of what it is like to be stung by a hornet, i.e. of the phenomenal character of a painful experience. This is what is left out by claiming that pain is C-fibres firing; it does not explain why C-fibres firing filling the functional role above should feel the way it does. As Levine puts it:

[16] Cf. §1.3.2. See Chapter 6 for a detailed discussion.

> "[T]here seems to be nothing about C-fiber firing which makes it naturally 'fit' the phenomenal properties of pain, any more than it would fit some other set of phenomenal properties. Unlike its functional role, the identification of the qualitative side of pain with C-fiber firing (or some property of C-fiber firing) leaves the connection between it and what we identify it with completely mysterious. One might say, it makes the way pain feels into merely a brute fact."[17] (Levine, 1983, p. 357)

Here what Levine relies on is the idea that when one explains a phenomenon by describing a causal mechanism then the fact that this mechanism is in operation must entail the presence of the phenomenon in question. If such a mechanism is in place but the occurrence of the phenomenon does not follow, then relying on the mechanism is inapt to explain the phenomenon (cf. Levine, 2001, p. 74). Identity statements in standard scientific explanations are explanatory exactly because they show how certain mechanisms are able to produce the very effects which are characteristic of the explanandum. So, for example, on the one hand, heat in classical thermodynamics is energy transferred from regions with higher temperature to regions with lower temperatures. On the other hand, molecules with higher mean kinetic energy collide more with other molecules in their vicinity thereby accelerating them, and thus transfer their kinetic energy to other regions occupied by molecules with lower original mean kinetic energy. 'Heat is mean molecular kinetic energy' is explanatory because once we understand the mechanism involving molecules we immediately see why heat is present.

The same cannot be said of identities concerning conscious experiences. No matter how detailed knowledge one has about the mechanisms of C-fibre firings one is still unable to see why the presence of feeling pain should follow. It seems to be equally conceivable that when one's C-fibres are firing one feels joy, or nothing at all. This is in stark contrast with the scientific cases: once we understand the properties and dynamics of H_2O molecules, or the mechanisms of momentum transfer we

[17] See Chapter 6 and Chapter 7 for more on brute identities.

immediately have a grasp why H_2O is water, or why mean molecular kinetic energy is heat, as opposed to something else.

In the standard scientific cases the identifications are made intelligible by a match between the causal-functional role of entities at the two sides of the identity. These functional roles fully characterise the entities in question. Contrary to this, conscious mental states cannot be fully characterised by their functional roles—over and above these functional roles there is their phenomenal character, the qualitative feel one experiences when undergoing them. This is why 'pain is C-fibres firing' is not fully explanatory: though it is explanatory with regard to the functional aspect of the concept pain, it is not explanatory with regard to its phenomenal aspect.[18]

Levine explicitly argues that from the fact that there is an explanatory gap no metaphysical conclusion follows, and thus the presence of the explanatory gap is compatible with physicalism (Levine, 1983, 1993, 2001). However, David Chalmers thinks that with an extra premise expressing a link between epistemic and ontological gaps Levine's explanatory gap can be turned into an argument against physicalism about consciousness. This extra premise says that what cannot be physically explained is not itself physical (cf. Chalmers, 1996, 2003, 2010b).[19] The resulting Explanatory Gap Argument is a version of the epistemic arguments against physicalism about consciousness. It proceeds as follows:

Explanatory Gap Argument

(P1) Physical accounts explain only structure and function.

(P2) Explaining structure and function does not suffice to explain consciousness.

[18] Roughly, the argument concluding on the 'heat is mean molecular kinetic energy' claim proceeds via two premises. The first premise is *a priori* saying that heat is whatever plays the heat role. The second premise is *a posteriori* expressing the observation that mean molecular kinetic energy plays the heat role. The identity statement follows from these two premises. Cf. §6.3 for a much more detailed discussion of the role identities play in reductive explanation.

[19] Cf. §2.2.1 and especially §2.3.

(P3) What cannot be physically explained is not itself physical.

(C) Physicalism about consciousness is false. (Cf. Chalmers, 2010b, pp. 105-106)

As a version of the epistemic argument the Explanatory Gap Argument relies on the lack of epistemic entailment. The kind of epistemic entailment in question here is explainability, and the epistemic gap is the explanatory gap—the observation that phenomenal truths cannot be explained by citing physical truths. To be more precise, the Explanatory Gap Argument claims that the implication from physical truths to phenomenal truths would require a functional analysis of consciousness. Since a functional analysis of consciousness cannot be given phenomenal truths are not implied by physical truths. This is established by the first two premises of the Explanatory Gap Argument. The extra third premise connects the epistemic conclusion to an ontological conclusion: it infers from the failure of explaining consciousness in physical terms to the conclusion that consciousness is not physical. From this the claim that physicalism about consciousness is false follows (Chalmers, 2003).

2.2.3 Chalmers' conceivability argument

According to Levine, the problem of other minds is one of the major symptoms of the fact that there is an explanatory gap. The problem is that no matter how detailed knowledge one might have about the physical processes taking place during one's conscious experience, one is unable to use this knowledge to decide whether other creatures are conscious or not. Since knowledge of the physical processes is knowledge of structure and function, and explaining structure and function does not tell us anything about the phenomenal character of conscious experience, we just simply cannot know whether the same physical processes are always accompanied by the same felt quality. Even if we accept it as an unexplained, brute fact that certain qualities are felt when particular physical processes take place, it only supports that creatures similar to us in their physical structure and functioning share the same phenomenal life we enjoy—it does not help with regard to creatures sufficiently different from us.

The so-called conceivability argument takes this line of thought a step further by emphasising that it is not necessary that when particular physical processes take place one always feel the same qualities. Since this connection is unexplained, we can always conceive of a situation where the physical processes are present but they are accompanied by no felt quality. The automaton above was an elementary example for this. The only requirement there was a similarity in functioning in a broad sense.[20] A more sophisticated example is of a total physical duplicate, whose physical structure and functioning down to the smallest detail is the same as ours. Such creatures (physical duplicates of humans), if they have no phenomenal life at all, are called *phenomenal zombies*.[21]

In contemporary literature David Chalmers is the main promoter of an argument against physicalism about consciousness starting from the premise that zombies are conceivable. He characterises zombies as follows:

> “So let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely. [...] To fix ideas, we can imagine that right now I am gazing out the window, experiencing some nice green sensations from seeing the trees outside, having pleasant taste experiences through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder. What is going on in my zombie twin? He is physically identical to me, and we may as well suppose that he is embedded in an identical environment, He will certainly be identical to me *functionally*: he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be *psychologically* identical to me

[20] Thus the automaton as it is presented in §2.1.2 is an example of the qualia-based arguments against *functionalism*. Cf. e.g. Block (1980).

[21] For early accounts of zombies see Campbell (1970) and Kirk (1974a, 1974b). Campbell talked about an ‘imitation man’, a physical duplicate of human beings who has no conscious experience at all. Kirk has coined the term ‘zombie’ for such a creature and argued that zombies were counterexamples to physicalism.

> [...]. He will be perceiving the trees outside, in the functional sense, and tasting the chocolate, in the psychological sense. All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions. He will even be 'conscious' in the functional senses described earlier—he will be awake, able to report the contents of his internal states, able to focus attention in various places, and so on. It is just that none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. There is nothing it is like to be a zombie." (Chalmers, 1996, p. 95)

The crucial question with regard to phenomenal zombies is not their natural possibility—Chalmers acknowledges that most probably there are no phenomenal zombies in the actual world, nor are they naturally possible (there are no zombies in those worlds which are like our world with respect to all laws of nature). What Chalmers is really interested in is whether they are metaphysically possible.[22] He argues that the best guide to whether phenomenal zombies are metaphysically possible is whether their notion is conceptually coherent. Chalmers claims that "if no reasonable analysis of the terms in question points toward a contradiction, or even makes the existence of a contradiction plausible, then there is a natural assumption in favor of logical possibility" (Chalmers, 1996, p. 96). If there was a conceptual entailment from physical processes to felt qualities then the notion of a phenomenal zombie would be conceptually incoherent. However, as the explanatory gap shows, there is no such conceptual entailment.

The conceptual coherence of phenomenal zombies might be attacked by claiming that those who think that phenomenal zombies are conceivable are, in fact, insufficiently reflective and thus overlook an incoherence (cf. Chalmers, 1996, p. 99). But the burden of proof in this case is on the opponent claiming that there is an incoherence lurking in the description of the situation. Without such proof, Chalmers

[22] In fact, Chalmers prefers the term 'logical necessity' over 'metaphysical necessity'. By logical necessity he means this: "B-properties supervene *logically* on A-properties if no two *logically possible* situations are identical with respect to their A-properties but distinct with respect to their B-properties" (Chalmers, 1996, p. 35) However, he acknowledges that his logical necessity is very close to metaphysical necessity. As he puts it: "the metaphysically possible worlds are just the logically possible worlds [... and the] metaphysical possibility of statements is logical possibility with an *a posteriori* semantic twist" (Chalmers, 1996, p. 38). Here I follow the vast majority of the physicalism literature by using the 'metaphysical' notion.

argues, we have every right to conclude that the notion of phenomenal zombies is conceptually coherent, and that phenomenal zombies are in effect metaphysically possible. Since the metaphysical possibility of physical duplicates of humans lacking phenomenal consciousness entails that physical facts does not necessitate metaphysically phenomenal facts, the falsity of physicalism about consciousness follows. That is, the conceivability argument has the following structure:

Conceivability Argument

(P1) It is conceivable that there are zombies.

(P2) If it is conceivable that there are zombies, it is metaphysically possible that there are zombies.

(P3) If it is metaphysically possible that there are zombies, then consciousness is non-physical.

(C) Physicalism about consciousness is false. (Cf. Chalmers, 2010a, pp. 106-108)

That is, the kind of epistemic entailment in question here is reflective conceivability[23], and the manifestation of the epistemic gap is the conceivability of the total absence of phenomenal consciousness even if the physical if fully present. The Conceivability Argument claims that the implication from physical truths to phenomenal truths would require that one cannot rationally conceive of all the physical truths without phenomenal truths. Since the thought of a phenomenal zombie seems conceptually coherent, i.e. one is able to rationally conceive of all physical truths without phenomenal truths, phenomenal truths are not implied by physical truths. This is established by the first premise of the Conceivability Argument. The second premise claims that there is a direct link between epistemic conclusions and ontological conclusions: it infers from the conceivability of zombies to the metaphysical possibility of zombies. The third premise expresses that if phenomenal zombies are metaphysically possible then the phenomenal character of

[23] I.e. considering if philosophical zombies can be ruled out on the basis of ideal rational reflection given all the physical truths. Cf. Footnote 13 in §2.2.1.

conscious experience is something over-and-above the physical. From this the claim that physicalism about consciousness is false follows.

2.2.4 Jackson's knowledge argument

The third influential argument against physicalism about consciousness is Frank Jackson's Knowledge Arguments (Jackson, 1982, 1986). Its basis is a thought experiment about Mary, a future scientist with a little deficit in experiences. As Jackson describes:

> "Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. If physicalism is true, she knows all there is to know. For to suppose otherwise is to suppose that there is more to know than every physical fact, and that is what physicalism denies. Physicalism is not the noncontroversial thesis that the actual world is largely physical, but the challenging thesis that it is entirely physical. This is why physicalists must hold that complete physical knowledge is complete knowledge simpliciter. […] It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. […] Hence, physicalism is false. This is the Knowledge Argument against physicalism in one of its manifestations." (Jackson, 1986, p. 291).

That is, Mary is a future scientist, who knows everything about the physical processes of colour vision. She has a complete knowledge in physical terms about what goes on in human brains when we see colours. However, apart from all this, Mary has never ever experienced colours. She grew up in a black-and-white room. She acquired all her knowledge via black-and-white books and black-and-white televisions. Then one day Mary is let out of her room and sees a ripe tomato. This is the first time she becomes acquainted with the experience of seeing something red. At this point the question is if the knowledge she acquires in this situation is a new piece of knowledge, or something she already knew from her previous studies.

Jackson claims that Mary does learn something new, something she did not already know: she learns what it is like to have a reddish experience. If so, argues Jackson, then physicalism is false, since before she came out of her room, Mary already knew all the physical facts about seeing something red, and such a complete physical knowledge should have entailed all the facts about seeing something red.

For Mary, who had a complete physical knowledge about red experiences, learning something new means that she learns something over and above the physical facts, i.e. that she learns about a further property of red experiences. This further property is the property of what the experience of seeing something red is like. It is a *phenomenal property*—a property not entailed by the full description of the physical facts and not identical with any physical-functional property.

The knowledge argument relies on two premises. The first one tells us that Mary has complete physical knowledge (about what happens when someone sees something red) before she leaves her black-and-white room; i.e. that she knows all the physical facts—everything physical there is to know—about having reddish experiences. The second premise tells us that Mary learns something new after her release, i.e. that she acquires some kind of new knowledge concerning facts about seeing something red —information she did not know before she left her room. From these two premises, the knowledge argument concludes that there are phenomenal (non-physical) facts about seeing something red, and thus physicalism must be rejected.[24] To put it more formally:

Knowledge Argument

(P1) Mary knows all the physical facts.

(P2) There are truths about consciousness that are not deducible from physical truths.

(P3) If there are truths about consciousness that are not deducible from physical truths, then there are non-physical facts.

(C) Physicalism about consciousness is false. (Cf. Chalmers, 2010a, pp. 108-109)

That is, the kind of epistemic entailment in question here is deducibility, and the epistemic gap appears as presented by the observation that phenomenal truths cannot be deduced from physical truths. The Knowledge Argument claims that the

[24] It might be worth noting that in the original version of the thought-experiment (Jackson, 1982) the knowledge argument is formulated using the notion of 'information'. The original argument runs as follows. P1: Mary has all the physical information about seeing something red without ever having seen red. P2: When she sees red for the first time Mary comes to know some information about seeing something red she did not know before. C: There is non-physical information about seeing something red; i.e. not all information is physical information. Cf. Nida-Rümelin (2010).

Horgan points out that this original version allows two interpretations. According to the weaker one 'having all the physical information' translates onto 'having complete physical knowledge', and according to the stronger one, onto 'knowing all the physical facts' (Horgan, 1984).

The weaker version can be put in the following way: (WeakP1) Mary has complete physical knowledge about seeing something red without ever having seen red. (WeakP2) When she sees red for the first time Mary comes to acquire some kind of knowledge concerning facts about seeing something red—something she did not know before. (WeakC) There is some kind of non-physical knowledge concerning facts about seeing something red; i.e. not all knowledge is physical knowledge. Note that this weaker conclusion is compatible with physicalism. Cf. §3.1.

By contrast, the stronger version runs as follows: (StrongP1) Mary knows all the physical facts about seeing something red without ever having seen red. (StrongP2) When she sees red for the first time Mary comes to know some facts about seeing something red she did not know before. (StrongC) There are non-physical facts about seeing something red; i.e. not all facts are physical facts. Cf. Nida-Rümelin (2010).

The main difference between the weaker and the stronger version is that whereas the weaker one has an epistemological conclusion (i.e. not all knowledge is physical knowledge) the stronger version concludes to an ontological claim (i.e. not all facts are physical facts). Although Jackson's original account (Jackson, 1982) implies both versions, from his later discussions (Jackson, 1986, 1998) it turns out that it is the stronger version what he has in mind.

implication from physical truths to phenomenal truths would require the deducibility of phenomenal truths from physical truths. Since it is shown that a perfect reasoner who knows all the physical truths is unable to deduce phenomenal truths from physical truths, phenomenal truths are not implied by physical truths. This is established by the first two premises of the Knowledge Argument. Again, the third premise connects the epistemic and ontological domains: it infers from the failure of deducing truths about consciousness from physical truths to the conclusion that there are non-physical facts (phenomenal properties instantiated by certain particulars). From this the claim that physicalism about consciousness is false follows.

2.3 The Entailment Thesis and *A Priori* Physicalism

In the previous sections we have seen the major versions of the epistemic argument against physicalism. In this final section of Chapter 2, I would like to concentrate on the main motivation behind these arguments.

One way of illustrating the very idea behind the supervenience-based formulation of physicalism[25] is this: two possible worlds cannot differ in their prima facie non-physical properties or laws without there being a difference in their physical properties or laws. To put it a bit more formally, physicalism is true at a possible world *w* if and only if any world which is a minimal physical duplicate of *w* is a duplicate of *w simpliciter* (cf. Jackson, 1994, p. 27).

Here a physical duplicate of *w* duplicates all the physical facts as well as all the physical laws (cf. §1.6.1). The ‘minimal’ clause is required to evade the problem of ‘epiphenomenal ectoplasm’, i.e. the case where there is a non-physical property which has no causal connections to anything else. Without this extra clause, physicalism would rule out such a world (cf. Horgan (1983) and Lewis (1983b)). Jackson’s proposal solves the problem by restricting the definition to minimal

[25] Saying, according to its initial formulation, that all prima facie non-physical facts metaphysically supervene on physical facts. Cf. §1.3, but also §1.6.4 for a ‘relativised’ formulation.

physical duplicates, i.e. to duplicate worlds which are identical with *w* in all physical respects but do not contain anything else.[26]

Physicalism, this formulation says, is true in the actual world, if copying everything physical from the actual world (and initially including nothing else in the copy-world) results in a copy-world, which is identical with the actual world in all respect. That is, from physicalism it follows that the physical facts (and laws) *entail* all the facts (and laws). Frank Jackson calls this consequence of physicalism the *Entailment Thesis*: if *ΣP* is the full physical description of a world *w* (describing all the physical facts and laws of *w*, and just them) and *Q* is an arbitrary (but true) fact of *w*, then, necessarily, *if ΣP then Q* (cf. Jackson, 1994).[27] In other words, if physicalism is true, then the material conditional 'if *ΣP* then *Q*' is a necessary truth.

Now given that this material condition is necessary, the question arises whether it is *a priori*. Can the material condition 'if *ΣP* then *Q*' be known *a priori*, or only just *a posteriori*? That is, does the full physical description of the world *a priori* entail such facts as, for example, the chemical fact that 'water boils at 212 °F at sea level', or the geographical fact that 'water covers 60% of planet Earth', or the phenomenal fact that 'this is what it is like to taste water'?

Proponents of the epistemic arguments against physicalism answer with a definite yes. They argue that the correct understanding of the entailment from all the physical facts to any arbitrary fact is *a priori*. That is, they argue that the correct understanding of physicalism is *a priori physicalism*, and the correct understanding of the Entailment Thesis is the *A Priori Entailment Thesis*: the physical facts *a priori*

[26] There are other problems with the correct formulation of physicalism which are well beyond the scope of this section. See e.g. Kim (1993a) for the 'lone ammonium molecule' problem, Jackson (1998) for the 'necessary beings' problem, and Hawthorne (2002) for the 'blockers' problem.

[27] Cf. 6.3.1, especially Footnote 18 there. David Chalmers argues that the correct formulation of the Entailment Thesis interprets *Q* as an arbitrary *positive* fact. As Chalmers puts it: "Negative existential facts such as 'There are no angels' are not strictly logically supervenient on the physical, but their nonsupervenience is quite compatible with materialism." (Chalmers, 1996, p. 41) This can be fixed either by restricting *Q* to positive facts or by adding an extra second order 'that's all' fact to the supervenience base of physicalism (cf. Chalmers, 1996, p. 41).

entail all the facts. In other words, the material condition that 'if *ΣP* then, e.g. water covers 60% of planet Earth' is *a priori* knowable. Once one knows all the physical facts about the world one is able to deduce any arbitrary fact about the world solely by means of *a priori* reasoning, without leaving the armchair.[28]

Given all this, it is now clear why establishing an epistemic gap between physical facts and phenomenal facts is so important for the arguments introduced in this chapter. For establishing an epistemic gap amounts to showing that there is no epistemic entailment from physical knowledge to phenomenal knowledge, i.e. that even if one knows all the physical facts one is still unable to deduce phenomenal facts by means of *a priori* reasoning alone. Now if physicalism entails the *A Priori Entailment Thesis* then the mere observation that phenomenal facts are not *a priori* entailed by the physical facts falsifies physicalism.[29]

[28] See Chapter 6 for an extensive discussion of the role this so-called '*a priori* passage' view (cf. Jackson, 2003) plays in reductive explanations.

[29] Or to be more precise, physicalism about consciousness. Cf. §1.6.4.

Chapter 3:
Phenomenal Concept Strategy

3.1 Introducing Phenomenal Concepts

Recall Jackson's Knowledge Argument. It starts with the claim that pre-release Mary knows everything there is to know about colour vision that can be learned in a black-and-white room (i.e. knows all the physical facts). Then it proceeds with the claim that post-release Mary learns something new, she learns a new fact she had not previously known (namely what it is like to see something red). From these two premises the Knowledge Arguments concludes that physicalism is false.[1]

If one wants to resist the conclusion of such an argument one can follow two strategies: either deny the soundness of the argument by claiming that one or both of the premises are false, or deny the validity of the argument by claiming that the conclusion does not really follow from the premises. Physicalists defending physicalism against the Knowledge Argument have tried all of these possibilities.[2] In this chapter I shall focus on the latest attempt to block the Knowledge Argument. This attempt is called Phenomenal Concept Strategy[3], and it is widely considered as a general strategy defending physicalism against all versions of the epistemic argument. In what follows, first I shall introduce the main idea behind the strategy itself in the context of the Knowledge Argument, and then focus on how phenomenal concepts—the key ingredients of the strategy—might be best understood.

[1] Cf. §2.2.4, and especially Footnote 24 there.

[2] For example, Churchland (1985) and Harman (1990) challenge the first premise by denying that pre-release Mary knows all the physical facts. Dennett (1991) challenges the second premise by denying that post-release Mary learns anything new. Lewis (1983a, 1990) and Nemirow (1980, 1990) follow the second route, and challenge the validity of the argument by claiming that the Knowledge Argument equivocates on the term 'knowledge': whereas pre-release Mary knows everything in terms of propositional knowledge (knowing that), post-release Mary learns something new in terms of abilities (knowing how).

[3] This term has been coined by Stoljar (2005).

3.1.1 There is something about Mary

Instead of the original Knowledge Argument, consider the following, slightly modified thought experiment. Suppose that Mary is not immediately released out into the 'wild' where she can freely encounter with all different kinds of coloured objects, but first is shown a piece of paper with a patch of red on it (cf. Nida-Rümelin, 1996). Now with this clever move, the experimenters can avoid cases where post-release Mary is able to infer the colour she is actually seeing on the basis of her shape recognition capabilities ('that is a London-bus') and her previous knowledge ('London-buses are red'). There is no way for Mary to infer the colour of the patch on the paper in a similar way. That is, Mary, in this new version of the thought experiment, is unable to refer to her colour experience with an old concept she acquired previously back in her room. However, she is nevertheless able to form propositions like "I will have *this* experience again before the day is out" (Papineau, 2002, p. 62). Propositions like this have truth-values and thus, by formulating such propositions, Mary, in fact, expresses propositional knowledge.

Note two things. First, since Mary in the scenario above is not able to use her old colour concept she acquired in her room to refer to her actual experience she is deploying new concepts, new vehicles of thought to think about her actual experience. The term '*this* experience' above is such a new vehicle of thought, a so-called *phenomenal concept*, which picks out Mary's actual experience. Second, by acquiring these new phenomenal concepts and utilising them, Mary becomes able to form novel proposition. That is, Mary gains new propositional knowledge.

In other words, post-release Mary does learn something new: a novel way of thinking. However, even if this is more than just a mere ability since it results in new propositional knowledge,[4] it does not necessarily mean that Mary learns something about a *new fact.* On the contrary, physicalists retort: she learns a new way of thinking about an old fact—a fact she has already learned inside her room in terms of

[4] Cf. the so-called Ability Hypothesis of Lewis (1983a, 1990) and Nemirow (1980, 1990) claiming that all that Mary learns are new routes to her old physical-functional concepts. See Footnote 2 above.

physical-functional concepts (e.g. terms like 'neural activation' etc.). That is, this line of thought argues, the Knowledge Argument is right in claiming that post-release Mary gains new propositional knowledge, however, it is wrong in moving from this claim to the further claim that this new propositional knowledge is knowledge of a new fact.[5]

3.1.2 Phenomenal Concept Strategy

Acknowledging that there are phenomenal concepts—special vehicles of thought with the aid of which one can think about one's own experiences in a way, which is distinct from thinking about these experiences in terms of physical-functional concepts—opens up the possibility of blocking the Knowledge Argument in its anti-physicalist form. Moreover, the same strategy is useful to block all versions of the epistemic argument against physicalism.

As we have seen, the epistemic argument against physicalism relies on an *a priori* interpretation of the Entailment Thesis: physical truths should *a priori* entail all the truths (cf. §2.3). The Explanatory Gap Argument, the Conceivability Argument, and the Knowledge Argument all start with establishing an epistemic gap between physical and phenomenal knowledge. Since according to the *A Priori* Entailment Thesis, the entailment from physical truths to phenomenal truths should be *a priori* knowable, the existence of such an epistemic gap straightforwardly leads to the denial of physicalism.

With phenomenal concepts in hand, however, it becomes possible to block this line of thought. Proponents of Phenomenal Concept Strategy argue that the existence of the general epistemic gap is due to a *conceptual gap* between phenomenal and physical-functional concepts. All that the first premises of the different versions of

[5] Cf. Footnote 24 of §2.2.4—this line of thought accepts the weak and rejects the strong interpretation of the Knowledge Argument.

the epistemic argument really show is just that there is this conceptual gap, i.e. that phenomenal concepts are conceptually irreducible to physical-functional concepts.

Physical-functional concepts are the concepts featuring in scientific terminology—they are objective, and pick out their referents from a third-person perspective, independently of one's own experiences. They typically refer by causal, functional, or structural descriptions. All the knowledge Mary could acquire in her black-and-white room were conveyed by such physical-functional concepts. Contrary to this, phenomenal concepts are accessible only via experiences. As we have seen, they are the vehicles of thought when one thinks about one's own experiences.

The fundamental tenet of Phenomenal Concept Strategy is the claim that there is a conceptual gap between phenomenal and physical-functional concepts. For this to be the case, proponents of the Strategy must characterise phenomenal concepts in a way which renders them conceptually irreducible to physical-functional concepts.[6] Typically they do so by emphasising that phenomenal concepts refer *directly*, i.e. they do not refer by descriptions but rather pick out their referents without relying on contingent modes of presentations.[7]

Pre-release Mary, since she has learned everything she knows via descriptions, could not acquire phenomenal concepts about colour experiences. She acquires them upon her first encounter with coloured objects, and thereby she acquires new propositional knowledge. Since her phenomenal concepts are conceptually irreducible to physical-functional concepts, there was no way she could conclude on this new piece of propositional knowledge before her release. This is the reason why Mary learns something new after her release. However, the new concepts post-release Mary acquires need not necessarily refer to new facts. It very well might be the case (and

[6] That is, according to Phenomenal Concept Strategy, phenomenal concepts neither *a priori* imply nor are *a priori* implied by physical-functional concepts.

[7] See more on this later. The rest of this chapter is devoted to understanding the special characteristics of phenomenal concepts.

proponents of the Phenomenal Concept Strategy argue that actually it is the case) that these new concepts refer to facts which has already been picked out by some of Mary's old physical-functional concepts. With this move, phenomenal concept strategy is able to block the second premise of the epistemic arguments: there really is an epistemic gap between phenomenal and physical knowledge, however, it is due to a conceptual gap and thus does not indicate the existence of an ontological gap.

3.1.3 *A posteriori* physicalism

Phenomenal Concept Strategy does not provide an argument showing how phenomenal and physical-functional concepts co-refer. It only draws attention to that there is nothing in the anti-physicalist arguments which would exclude this possibility. In this sense, Phenomenal Concept Strategy is a *defence* of physicalism—it does not argue for physicalism, but rather shows that the presence of an epistemic gap is not necessarily fatal for physicalism. Phenomenal and physical-functional concepts, even if they are conceptually irreducible to each other, could nevertheless refer to the same property. That is, according to Phenomenal Concept Strategy, ontological conclusions do no follow from epistemic considerations concerning phenomenal knowledge.

The version of physicalism Phenomenal Concept Strategy defends is *a posteriori* physicalism. It acknowledges that the Entailment Thesis holds—physical facts entail all the facts,—however, denies that the entailment is *a priori* knowable. The move from physical knowledge to phenomenal knowledge is only *a posteriori*: one cannot infer phenomenal truths from physical knowledge by means of *a priori* reasoning alone, but needs to rely on further experiences. Thereby, the typical examples of how phenomenal and physical-functional concepts co-refer are the classical cases of Kripkean *a posteriori* identities, like 'Cicero is Tully', or 'water is H_2O' (Kripke, 1980).[8]

[8] See Part III (Chapter 6 and Chapter 7) for an extensive discussion of *a priori* and *a posteriori* entailment, and the role of identities in reductive explanations in general.

3.2 The Locus Classicus: Brian Loar's Account

In this and the forthcoming sections my aim is to clarify how different proponents of the Phenomenal Concept Strategy characterise phenomenal concepts. I start with Brian Loar, who proposed the initial version of Phenomenal Concept Strategy (Loar, 1990, 1997). Loar's main motivation was to put forward an account of phenomenal concepts which could explain how phenomenal concepts can pick out a physical property via a non-contingent mode of presentation, and be conceptually independent of all physical-functional concepts. According to his solution, phenomenal concepts are direct recognitional concepts.

3.2.1 Recognitional concepts

According to Loar, so-called recognitional concepts "have the form 'x is one of *that* kind'; they are type-demonstratives. These type-demonstratives are grounded in dispositions to classify, by way of perceptual discriminations, certain objects, events, situations." (Loar, 1997, p. 600)

So, for example, when an expert chicken sexer decides of a particular chick whether it is a male or a female, she employs a recognitional concept. Untrained individuals lack the crucial recognitional concepts—to them, male and female day-old chicks look similar. By being able to discriminate male- and female cloacae one becomes able to bring certain chicks under one concept (e.g. 'this one is of the male-type'), and other chicks under another concept (e.g. 'this one is of the female-type'). Note that the names 'male' and 'female' as denoting the two kinds are insignificant here. The recognitional concepts one acquires during a chicken-sexer training allows one to recognise two sets of the stimuli (cloacae) as two different types regardless whether one is informed about the public language names of these types.[9]

[9] Sure one must have two names to refer to these two types one becomes able to discriminate, however, their role is only to anchor the two types at the level of public language. Cf. §3.5 for a related account, namely, David Papineau's recent view on perceptual concepts.

Loar emphasises four characteristics of recognitional concepts. First, a recognitional concept is “recognitional at its core” (Loar, 1997, p. 601). That is, one cannot extract recognitional concepts from descriptions; one acquires them only via perceptual categorisation. In other words, possessing a recognitional concept requires the exercising of a recognitional ability.[10]

Second, recognitional concepts “need involve no reference to a past instance, or have the form ‘is of the same type as that (remembered) one’” (Loar, 1997, p. 601). That is, bringing an actual perceptual stimulus under a certain recognitional concept does not necessarily involve active comparison—one is able to judge that the actual perceptual stimulus is ‘another one of *those*’ without necessarily comparing the actual stimulus to a particular past instance (cf. Loar, 1997, p. 601).[11]

Third, “recognitional abilities depend on no consciously accessible analysis into component features; they can be irreducibly gestalt” (Loar, 1997, p. 601). Just as expert chicken sexers are not really capable of producing a positive description of how a male cloaca looks like, recognitional concepts do not necessarily reveal the details of the perceptual stimuli.

Fourth, “recognitional concepts are perspectival; [... they are] in part individuated by their constitutive perspective” (Loar, 1997, p. 601). That is, perceiving the same object from different perspectives might lead for someone to form one recognitional concept from one perspective and another from another perspective, thus failing to recognise the two sets of stimuli as presenting the same object (since the two recognitional concepts are *a priori* independent).

[10] Recognitional ability might be characterised as a capacity to store perceptual templates, compare actual perceptual stimuli against these templates, and judge them as similar or different. Cf. §3.5 for a more detailed discussion.

[11] Again, see §3.5 for an account detailing how stored templates can abstract away from particular instances.

3.2.2 Conceptual independence

Loar’s proposal is that phenomenal concepts are recognitional concepts picking out certain physical properties of the brain. Phenomenal concepts are those recognitional concepts which we deploy in phenomenological reflection: they reveal certain properties of the brain phenomenologically—independently of the physical-functional descriptions of these brain properties. As Loar formulates the fundamental tenet of Phenomenal Concept Strategy (cf. §3.1.2): “phenomenal concepts are conceptually independent of physical-functional descriptions, and yet pairs of such concepts may converge on, pick out, the same properties.” (Loar, 1997, p. 602).

Of course, the conceptual independence of phenomenal concepts from physical-functional descriptions plays a crucial part in Loar’s account: it explains why no *a priori* entailment follows from the fact that they both pick out the same property. Justifying the claim of conceptual independence, then, is of fundamental importance for Loar’s account.

Loar supports the conceptual independence claim by appealing to the recognitional nature of phenomenal concepts. He says that “recognitional concepts and theoretical concepts are in general conceptually independent” (Loar, 1997, p. 602). Theoretical concepts (i.e. physical-functional concepts) pick out their referents via analysing them in scientific terms, i.e. describing their causal and functional properties—what causal roles they play, and how their component parts form a certain structure. Contrary to this, recognitional concepts discriminate their referents without analysing them in scientific terms. As Loar puts it: “basic recognitional abilities do not depend on or get triggered by conscious scientific analysis” (Loar, 1997, p. 603).

Note, however, that this latter claim is not true in general, thus relying on the recognitional nature of phenomenal concepts will not suffice on its own. Though it is true of some examples that no amount of “book learnin’” (cf. Papineau, 2002) can help one to recognise first instances of certain experiences, it is definitely not true of all possible perceptual cases. On the one hand, it seems to be true of recognising a patch of red, for example, whereas, on the other hand, it seems to be false with regard to chicken sexers. As Biederman and Schiffrar (1987) showed, after reading a short description explaining how typical male and female cloacae look like (something very hard to learn by experience) inexperienced subjects were able to categorise problematic cases at the same level of accuracy as experienced chicken-sexers could do (whereas without reading such descriptions, the accuracy of untrained subjects' performance was significantly worse). What this tells us is that physical/theoretical descriptions are able to affect recognitional concepts.[12]

Though the third feature of recognitional concepts Loar cites[13] seems to be correct, it is not equivalent with the claim that descriptions of component features and recognitional concepts of the 'wholes' are always conceptually independent. It well might be the case—as it is, in the case of chicken sexers—that independence is true in one direction (no matter how experienced a chicken sexer is, it is extremely hard for her to formulate a positive description of how a male or female cloaca looks like), but not true in the opposite direction (information about component features, parts, and structures triggers better recognitional abilities).

To recap the ideas presented so far: phenomenal concepts are special recognitional concepts picking out physical properties (brain states), which are nevertheless conceptually independent of physical-functional descriptions of the very same brain states. Loar puts great effort into explaining the conceptual independence claim by relying on two further claims: first that phenomenal concepts refer directly, and second that the physical-functional properties, when picked out by phenomenal concepts, provide their own mode of presentation. In the next two sections I shall analyse these two claims respectively.

[12] Compare this with §4.1, where a similar observation plays a crucial role.

[13] "Recognitional abilities depend on no consciously accessible analysis into component features" (Loar, 1997, p. 601).

3.2.3 Direct reference

Loar says that phenomenal concepts are *direct* recognitional concepts. Phenomenal concepts are special recognitional concepts because they do not connote contingent modes of presentation that are metaphysically independent of the kinds they pick out. Referring directly, then, means picking out the referent without being mediated by contingent modes of presentation.[14] What the Direct Reference claim says, then, is this:

(DR) **Direct Reference:**
Phenomenal qualities, when picked out by phenomenal concepts, are picked out directly.

To put it in another way, phenomenal concepts and theoretical concepts have different conceptual roles. Theoretical concepts allude to scientific descriptions, and thus to contingent modes of presentation.[15] Phenomenal concepts, on the other hand, present their referents in a direct way. Grasping a phenomenal concept directly reveals the essence of the referent: that is, when exercising a recognitional concept one is in a position to know the essence of the referent.[16]

However, one might wonder why there is an epistemic gap between one's phenomenal knowledge employing phenomenal concepts and one's physical-functional knowledge employing fundamental scientific terms, if the referent of both ways of thinking is the same property. The reason why one might find the presence of the epistemic gap problematic is this. On the one hand, as it is advocated by Loar, phenomenal concepts reveal the essence of their referents. On the other hand, some theoretical terms also reveal the essence of the property they pick out: fundamental

[14] Note that having contingent modes of presentations does not exclude referring directly. Cf. Kaplan's content/character distinction (Kaplan, 1989). See also §3.3.2.

[15] Some of these scientific descriptions nevertheless pick out scientific essences. See below.

[16] The essence of a thing is "a property of it such that necessarily it has it and nothing else does" (Lewis, 1999, p. 328). Cf. also Stoljar (2009b).

scientific descriptions (if correct) capture the basic nature of the entities they talk about—there can be no 'more fundamental' entities playing e.g. the electron-role (given that electron-talk is part of the fundamental theory). That is, fundamental theoretical terms capture the essence of their referents in the sense that they are "conceptually interderivable with some theoretical predicate that reveals the internal structure of the designated property" (Loar, 1997, p. 603).[17]

What we have then, if both concepts refer to the same property, is two direct grasps on the essence of a certain property. Anti-physicalists argue that two such direct grasps need to reveal the essence transparently: one must be able to recognise that the two grasps reveal the same essence. However, this is not the case—grasping the essence of, say, pain via a directly referring phenomenal concept does not put us into an epistemic situation where we can immediately recognise that this essence is the same that can be grasped via a fundamental theoretical term; nor *vice versa*. Therefore, the anti-physicalist argument goes, the two concepts must refer to two distinct properties.

Loar claims that the expected transparency is only an illusion. He says: "what generates the problem is not appreciating that there can be two conceptually independent 'direct grasps' of a single essence, that is, grasping it demonstratively by experiencing it, and grasping it in theoretical terms" (Loar, 1997, p. 609). In other words, Loar claims that the transparency argument above equivocates on 'capturing the essence'—phenomenal concepts and fundamental theoretical concepts capture the essence of their referents in a different way (cf. Loar, 1997, p. 603). How fundamental theoretical terms reveal the essence of the referent seems to be uncontroversial. Thus understanding exactly how demonstrative grasping goes (what referring directly for a phenomenal concept amounts to and how it is performed) is a fundamental part of making sense of Loar's account.[18]

[17] For alternative views on fundamental physical properties consider Hawthorne's causal structuralism (2001), or Russellian monism (Russell, 1927; Stoljar, 2001).

[18] See §3.3 for more on clarifying how direct reference can be accounted for.

3.2.4 Own mode of presentation

In order to fully appreciate Loar's proposal, let's consider first why phenomenal concepts are *special* recognitional concepts. There is a trivial difference between a phenomenal concept and the recognitional concept an experienced chicken sexer utilises when recognising and categorising a particular cloaca as of the male-type: whereas the recognitional concept of this latter type picks out an external property (the male cloaca), phenomenal concepts pick out certain internal properties (brain states, as argued by physicalists).

It is not enough, however, to restrict recognitional concepts to those self-directed ones which pick out an internal property. One can deploy a self-directed recognitional concept to pick out certain bodily states other then phenomenal qualities. Loar's example is picking out cramps. Cramps are muscle contractions typically accompanied by a characteristic cramp-feeling. Let's consider the scenario when a cramp occurs in one's body and one experiences the characteristic cramp-feeling. In such a scenario, one is able to focus one's attention either to the cramp-feeling, or to the cramp itself. In the first case one picks out a particular brain state (given physicalism is true) by a phenomenal concept categorising it as a feeling of the cramp-feeling-type, whereas in the second case, one picks out the cramp itself by a self-directed recognitional concept categorising it as a particular muscle contraction at a particular site of one's body.

When one is deploying the cramp concept (the recognitional concept picking out muscle contraction), one is referring to "a muscular property indirectly, by way of a causal chain that is mediated by the phenomenal quality associated with the concept" (Loar, 1997, p. 604). That is, the cramp concept refers by way of a contingent mode of presentation: cramps cause cramp-feelings, the actual cramp-feeling focuses one's attention to the cramp itself, and one recognises the particular cramp as something of the cramp-type by deploying a self-directed recognitional concept (cramp concept).

Contrary to this, argues Loar, when one is deploying a cramp-feeling concept (a phenomenal concept picking out a brain state), one is referring to a brain property *directly*, not by way of a contingent mode of presentation. The phenomenal concept picks out the particular cramp-feeling (brain state) as one of the cramp-feeling type. Here the token cramp-feeling focuses one's attention to the phenomenal quality[19] type 'cramp-feeling'. As Loar puts it: "a phenomenal concept has as its mode of presentation the very phenomenal quality that it picks out" (Loar, 1997, p. 604). This is what I shall refer to from now on as the Own Mode of Presentation claim:

> (OMP) **Own Mode of Presentation:**
> Phenomenal qualities, when picked out by phenomenal concepts, provide their own modes of presentation.

The whole picture Loar presents, then, is this. By deploying a fundamental physical-functional concept (theoretical description) one picks out a certain physical (brain) property under a direct grasp which reveals the physical structure of the property. By deploying a phenomenal concept one is able to pick out the very same physical (brain) property under a direct but different grasp which reveals the essence of the property via a demonstrative act, mediated by a mode of presentation which involves the experience itself, resulting in recognising it as 'of *that* feeling-type'. Since two concepts with these characteristics are irreducible to each other no *a priori* connection can be drawn between the two ways of conceiving the same property.

Note that on Loar's account the OMP-claim is an unexplained explainer—Loar uses the OMP-claim to explain the DR-claim: he accounts for how phenomenal concepts refer directly by claiming that phenomenal qualities when picked out by phenomenal concepts provide their own mode of presentation. However, the OMP-claim itself is left quite unexplained. In what follows I shall introduce more accounts of

[19] I use the term 'phenomenal quality' neutrally, i.e. as a term *not* implying that there is any non-physical property involved.

phenomenal concepts and concentrate on their contribution to understanding Loar's Direct Reference and Own Mode of Presentation claims.

3.3 Elaborating on the Direct Reference Claim

First, let's consider two accounts of phenomenal concepts which claim to clarify Loar's Direct Reference claim. The first one is Michael Tye's causal-recognitional account (Tye, 2003), the second is the demonstrative/indexical account advocated by John Perry (2001), John O'Dea (2002), and Janet Levin (2008).

3.3.1 The causal-recognitional account

Tye (2003) subscribes to the view that appealing to special phenomenal concepts is the best way to defend physicalism about consciousness against anti-physicalist objections. Phenomenal concepts, for him, are special because of three reasons.

First, they are not physical concepts. Given that phenomenal concepts are the vehicles of thought one deploys when one becomes aware of the phenomenal character of one's experience, exercising a phenomenal concept means focusing attention on what it is like to have that particular experience.[20] Were phenomenal concepts physical concepts, Mary would know what it is like to see red well before her release, since she already knows everything there is to know about colour vision in physical terms. However, according to the conclusion of the Knowledge Argument, this is not the case: Mary acquires new knowing-that (though of an old fact) when she leaves her room and sees red for the first time. Since phenomenal concepts are not physical concepts, Mary does not possess them while she lives in her room—she acquires them only when she first sees something red and attends to the colour experience she is having.

[20] According to Tye, phenomenal concepts refer to experience types. See §3.5.2 for an argument to the very same conclusion.

Second, phenomenal concepts are not demonstrative concepts utilising physical sortals. Tye asks us to imagine a modified version of the Mary thought experiment. In this modified version Mary is able to bring another person's brain states under recognitional concepts by using a future device called cerebroscope. The cerebroscope connects the other person's brain with Mary's room in a way which makes Mary able to read from the device what particular brain state the person is in. So with the aid of this device Mary can form thoughts like 'that person is in brain state F' where 'F' is a physical predicate expressing the physical properties of the brain state. No matter whether that particular brain state picked out by Mary as 'brain state F' corresponds to the person's having the experience of seeing something red, it is still true that when Mary leaves the room and sees something red for the first time she makes a new discovery and acquires a new piece of knowledge.

Third, phenomenal concepts have no *a priori* associated co-referential physical concepts. This third claim says that someone who possesses a phenomenal concept cannot know *a priori* what physical concept picks out the same referent as the phenomenal concepts does. This is the very point the so-called Explanatory Gap Argument draws attention to: it is always intelligible to ask why a certain physical state feels the way it does (even if an identity theory is true). Co-referential phenomenal and physical concepts are conceptually independent. In other words, what this third feature of phenomenal concepts shows is that the reference fixers of phenomenal concepts are not *a priori* associated physical descriptions.

From all this, Tye argues, it follows that phenomenal concepts refer directly. It is not the case that phenomenal concepts refer via a contingent mode of presentation. There are no properties distinct from the referent and picked out by certain physical descriptions which could present the referent itself by being *a priori* associated to it. Phenomenal concepts have "no associated reference fixers", they have "no descriptive content at all" (Tye, 2003, p. 95). As Tye summarises his own account:

"phenomenal concepts are non-demonstrative, general concepts that refer directly without the assistance of any associated reference-fixers." (Tye, 2009, p. 51)

Tye accounts for direct reference in causal terms. As he puts it:

> "my proposal [...] is that phenomenal concepts refer via the causal connection they have with their referents. In first approximation, a phenomenal concept C refers to a phenomenal quality Q via C's being the concept that is exercised in an introspective act of awareness by person P if, and only if, under normal conditions of introspection, Q is tokened in P's current experience and because Q is tokened." (Tye, 2003, p. 97)

That is, what Tye proposes is a causal covariation account: phenomenal concepts and their referents—phenomenal qualities—causally covary with each other. Phenomenal concepts refer directly due to there being a causal link between them and their referents. When one focuses one's attention toward an experience one's having a certain phenomenal concept gets exercised *because* a particular phenomenal quality is tokened in one's experience. Had a different type of phenomenal quality been tokened, a different phenomenal concept would have been exercised.

This cannot be the whole story, however. A pure covariation-based formulation of causation is not sufficient, since effects of common causes also covary with each other despite the fact that they themselves are not causally linked. What is needed to overcome this pitfall is an asymmetric dependence condition: if a state S causally covaries under normal conditions with feature F and also with some other feature G then S and F are causally linked under normal conditions only if "were F to fail to covary with G, the causal covariation link between S and F under optimal (normal) conditions would still hold but that between S and G would be broken" (Tye, 2003, p. 98). This further condition solves the problem posed by cases where a phenomenal concept covaries with both a phenomenal quality and some other physical property which itself is not a phenomenal quality. The phenomenal concept refers to the phenomenal quality as long as were the phenomenal quality to fail to covary with the non-phenomenal quality, the phenomenal concept would still causally covary with the phenomenal quality but not with the non-phenomenal quality.

This is still not the whole picture yet. For one might object that a more developed version of a cerebroscope poses a problem. Remember, in a previous example, a cerebroscope connected another person's brain to Mary's room in a way that Mary could read (say from a screen) what brain state the person was in. Now imagine that the cerebroscope is linked directly to Mary's brain (thus connecting the other person's brain to Mary's brain) in a way which makes it possible for Mary to recognise the other person's brain states by the act of introspection. Imagine next that the other person sees something red. Due to being wired to the other person's brain in the appropriate way, Mary now is able to recognise the other person's brain state (which *is* the distinctive phenomenal character of seeing something red) and think that the other person is seeing something red without Mary herself actually experiencing seeing something red.

This case fulfils all the requirements of the causal covariation account. Mary acquires a concept which refers directly to the phenomenal quality of seeing something red, but this concept is not a *phenomenal* concept since Mary still doesn't know what it is like to see something red.[21] That is, not all concepts that refer directly are phenomenal concepts. 'What makes a concept refer directly' is a different question than 'what makes a concept phenomenal'. Tye's proposal for answering the former question is the causal covariation account. Nonetheless, he needs to add further supplement in order to answer the latter question.

[21] Tye's original example goes like this: "Fred is a 21st century neuroscientist who is incapable himself of feeling pain in virtue of a neurological defect he has had since birth. Fred has a device partly wired into his brain that causes him to think that another person is feeling pain when and only when the external part of the device is directed at the other person's brain and the relevant brain state is present there. Fred's thought exercises a concept of pain, but that concept isn't a phenomenal concept. For Fred does not know what it is like to experience pain, and intuitively one cannot grasp the phenomenal character of pain, one cannot have a *phenomenal concept* of pain, without knowing what it is like." (Tye, 2003, p. 98)

Tye thinks that what makes a concept that refers directly to a phenomenal quality a *phenomenal* concept is that "it functions in the right sort of way" (Tye, 2003, p. 99). Due to the conceptual irreducibility of phenomenal concepts, however, this functioning cannot be specified *a priori* in purely non-phenomenal terms. The best one can get, Tye argues, is the following specification:

> "A concept is phenomenal […] if and only if (1) it is laid down in memory as a result of undergoing the appropriate experiences [...] (2) it tends to trigger appropriate conscious images (or quasi-images) in response to certain cognitive tasks, and (3) it enables its possessors to discriminate the phenomenal quality to which it refers directly and immediately via introspection." (Tye, 2003, p. 99)

Tye also ads that this functioning brings with it a distinctive first person perspective on phenomenal qualities. I shall elaborate on important aspects of this specification of what makes a phenomenal concept phenomenal in the subsequent sections where I shall compare other accounts of phenomenal concepts with Tye's approach.

The fundamental point now—the moral of this section—is this. Tye's causal-recognitional approach accounts for the direct reference of phenomenal concepts in terms of causal covariation. What makes a phenomenal concept directly refer to a phenomenal quality is that the phenomenal concept gets exercised only if the appropriate phenomenal quality is being tokened and because it is being tokened.

3.3.2 The demonstrative/indexical account

According to the so-called demonstrative/indexical account (Perry, 2001; O'Dea, 2002; Levin, 2007) phenomenal concepts work like demonstratives or pure indexicals. Indexicals, in general, are terms the referents of which are "dependent on the context of use" (Kaplan, 1989, p. 490). Demonstratives are special indexicals, which refer to an object picked out via an act of demonstration, i.e. a "presentation of a local object discriminated by a pointing" (Kaplan, 1989, p. 490)—e.g. 'this', 'that'. Pure indexicals are indexicals for which "no associated demonstration is required" (Kaplan, 1989, p. 491)—e.g. 'I', 'now', 'here'. That is, what a demonstrative or pure indexical refers to depends on the specific context of the actual use. My utterance 'that chick' might pick out as its referent one of the photos used in the classical Biederman and Schiffrar experiment (1987)[22], or a cute little plush animal my niece usually plays with depending on whether I am pointing out to the article, or the toy. Similarly, the utterance 'here' might pick out Edinburgh, or Budapest depending on whether the utterer at the time of the actual utterance is in Edinburgh, or in Budapest.

Utterances in general are uttered in a particular context and are evaluated in a particular circumstance. The context is a "possible occasion of use" (Kaplan, 1989, p. 494) of an utterance.[23] The circumstance of evaluation is a possible state of the world on the basis of which one tries to determine the truth of falsity of the utterance (cf. possible worlds in Kripke, 1980).[24]

Indexicals are special in at least two respects. First, as we have already seen, the context of utterance plays an important part in determining the content—and ultimately the reference—of certain, indexical involving, utterances. The utterance 'I submit my dissertation in Edinburgh' expresses the proposition that 'Peter submits his dissertation in Edinburgh' if it is uttered by myself, and is true if evaluated under actual circumstances, whereas it expresses the proposition that 'Gina submits her dissertation in Edinburgh' if it is uttered by my wife, and is false if evaluated under actual circumstances. 'I' picks out me in the first case, and my wife, Gina, in the second case. This is what Kaplan (1989) calls the *content* of an indexical: it is what a particular use picks out—the actual utterer in the case of 'I', the actual place of the utterance in the case of 'here', the object the utterer points at in the case of 'that', etc.

[22] Cf. §3.2.2.

[23] A context is characterised by an *agent* (the actual utterer), a *time* (the actual time of the utterance), and a *location* (the actual place of the utterance).

[24] By evaluating of a sentence under a given circumstance one gets a truth value, whereas by evaluating a singular term under a given circumstance one gets an object.

The *character* of an indexical is what determines “the content in varying contexts” (Kaplan, 1989, p. 505), i.e. defines how the context of use affects the content. That is, the character is a function from contexts to contents, it tells us how the content changes from context to context—e.g. it tells us that the content of ‘I’ is ‘the actual utterer’, whoever it is in the actual context, the content of ‘here’ is ‘the actual place of utterance’, wherever the utterer is in the actual context, etc. (cf. Speaks, 2011).[25]

Second, the reference of indexicals depends only on the context of use, and not on the circumstance of evaluation. Once one determines the referent of an indexical in a particular context, it remains the same under different circumstances. Contrary to this, the reference of a non-indexical term is determined by the circumstance of evaluation and not the context of use. Compare the term ‘that chick’ with the term ‘the famous chick featuring in the Biederman-Schiffrar article’. Imagine that I utter the first sentence while pointing at Figure 2 in Biederman and Schiffrar’s article. In this case, the content of the first utterance is ‘the chick Peter is pointing at on the 3rd of September, 2011 while sitting in his living room in Budapest’, whereas the content of the second utterance is ‘the chick with the property of being utilised as an illustration in the Biederman-Schiffrar article’. Now, evaluated under the actual circumstance, both terms pick out the same object: the particular day-old-chick the picture of which has been used as an illustration in the 1987 Journal of Experimental Psychology paper co-authored by Biederman and Schiffrar. However, if I evaluated the two terms in a possible world where Biederman and Schiffrar, instead of studying chicken-sexer, went on to study, say, children’s categorisation of plush animals and artefacts, and utilised a photo of a plush chick as an illustration, then the references of the two terms would come apart—the first term would still pick out the same day-old-chick, whereas the second term would refer to the plush chick.

[25] In the case of non-indexicals the content does not change with different contexts. What determines the content is the content itself, or as it is often put, the character of non-indexicals coincides with their content.

Unlike the description utilised in the second term, the indexical ‘that’ fixes the reference of ‘that chick’ regardless the possible circumstances of evaluation. The second term refers via a description, i.e. picks out its referent via contingent properties which can change from circumstance to circumstance, whereas the first term, the indexical, refers *directly*. As Kaplan puts it, the semantic rules of indexicals “provide *directly* that the referent in all possible circumstances is fixed to be the actual referent” (Kaplan, 1989, p. 493).

Proponents of the demonstrative/indexical account of phenomenal concepts capitalise on these features of indexicals. For example, John O’Dea (2002) argues that similarly to how the character of indexicals determine what they pick out depending on the context, there is a similar ‘rule’ for phenomenal concepts determining their referents. So, for example, just as the referent of, say, ‘here’ (as in ‘I am here’) is determined by the rule ‘the actual location of the utterance’, the referent of, say, ‘pain’ (as in ‘I am in pain’) is determined by the rule ‘the kind of state the utterer is actually in when she injures herself’ (cf. O'Dea, 2002, p. 175). Note that ‘pain’ here is a public language word. O’Dea argues, that we learn to use the word ‘pain’ in situations when we injure ourselves to describe the sensation we are having, and thereby we acquire a phenomenal use of this word.[26] Now this phenomenal use of ‘pain’ will pick out the kind of state the user is actually in when she uses the term regardless whether the user is a member of the normal population (and thus is having an experience similar to the painful sensation we typically have in a similar situation) or, say, is an invert and is feeling pleasure every time she

[26] Similarly, O’Dea argues, the way we learn the names of colour sensations—a phenomenal use—is “parasitic on the way we learn the names of *colours*” (O'Dea, 2002, p. 174)—a perceptual use (cf. §3.5). Typically, a child learns such words in situations where, for example, her parent points to the grass and says ‘that colour is green’. The child thereby learns the name of the colour of grass (whatever phenomenal quality she experiences when she looks at the grass), and consecutively she learns to refer to her phenomenal state by an inward ostension ‘that experience is green’ (cf. O'Dea, 2002, p. 174).

injures herself.[27] In this respect, phenomenal concepts work like indexicals do: the kind of state they refer to is jointly determined by their character and the actual context of their use (i.e. who uses them). The character of 'pain' is 'the kind of state the utterer is actually in when she injures herself' both for the invert and the non-invert—it is invariant to different uses of the term. However, the content of 'pain' is different for the invert (state X; i.e. normal pleasure state) and for the non-invert (state Y; i.e. normal pain state)—and it might very well change for every user.[28]

Note that the emphasis, here, is *not* on public language terms like 'pain' or 'red', not even on *their* phenomenal use. The same story might be told with demonstratives as well. Consider Nida-Rümelin's version of the Mary thought experiment (cf. §3.1.1). In that case, when Mary is finally let out of her room she is shown a sheet of paper with a patch of red on it. In this case, lacking any clues which could be matched up with the rules she has learned in her room about the use of public language terms, she is unable to use the term 'red'. Still, she can refer to her particular experience by using the term 'that experience'. Just as one might use the term 'that chick' together with an outward ostension, one uses the term 'that experience' together with an *inward* ostension to pick out a particular experience one is having. The character of 'that experience' is something like 'the kind of state the utterer is actually introspectively pointing at', whereas its content is the actual phenomenal state itself, which, again might change for every token use. Nonetheless, once the content is fixed by an actual use, the referent won't change no matter under what circumstance one tries to evaluate the proposition containing the phenomenal concept in question.

[27] An *invert*, in general, is a person who is similar to us in every physical respect, nevertheless, certain experiences she is having are inverted relative to those we are having in similar situation. So, for example, a 'colour-invert' when looking at ripe tomatoes has experiences similar to those we have when looking at the grass, and *vice versa*. Similarly, a 'pain-invert' might be someone, who feels pleasure in situation we normally feel pain, and *vice versa*.

[28] Cf. O'Dea (2002, pp. 177-178). Note that states X and Y are private states. This, however, isn't essential to the demonstrative/indexical account. Indexicals typically pick out accessible features of the context of utterance. Nevertheless, the content of indexicals is irreducible to any descriptive content. This is why, so the argument goes, phenomenal concepts are irreducible to non-phenomenal concepts.

That is, phenomenal concepts refer directly—as opposed to via contingent properties—in the same sense indexicals do.

Levin (2007) goes a step further in characterising direct reference. She draws attention to the fact that though the character of indexicals map contexts to contents and thus determine the content of the indexical, this content is irreducible to the character or to any other descriptive content. That is, indexicals and phenomenal concepts alike, do not refer via reference fixing descriptive modes of presentation, but rather refer directly. This direct reference, Levin argues, might best be characterised by shifting focus from pure indexicals to demonstratives. Type-demonstratives are mere pointers directed at certain types[29] of experience. Where the pointer is directed at is determined by which property causes differentially the use of the particular demonstrative. In line with this, the experience causally responsible for the application of a certain phenomenal concept in the introspective act of recognition and re-identification is the referent of that particular phenomenal concept. This is how phenomenal concepts refer directly. As Levin puts it, the referent of phenomenal concepts is:

> "determined solely by the causal and dispositional relations an individual has to her internal states that are effected by an introspective 'pointing in'; that is, by the *fact* that she's in causal contact with a certain property and is disposed to reidentify it on subsequent occasions." (Levin, 2007, p. 89, original emphasis)

When Nida-Rümelin's version of Mary is shown a piece of paper with a patch of red on it, she immediately becomes able to subsequently recognise and re-identify that reddish experience as another instance of 'that experience'. Upon a novel encounter, she would be able to express the proposition that 'I have already had that experience before'. 'That experience' here refer to the reddish sensation Mary is having, argues Levin, because this sensation is that triggers—or, as Levin puts it, differentially

[29] Levin argues that it is the class of phenomenal concepts picking out experience types (as opposed to tokens) which are of crucial importance since only they can play a part in the recognition and re-identification of an experience. Cf. §.3.5.

causes (cf. Levin, 2007, p. 91)—the application of the phenomenal concept ('that experience') in terms of which Mary thinks about the sensation.

As a moral, note that the demonstrative/indexical approach subscribes to a very similar causal account of direct reference than the causal-recognitional account does. This causal account of reference is a common ground in most contemporary approaches to phenomenal concepts. However, some argue, much more need to be said about the relation between phenomenal concepts and their referents than the postulation of a brute causal link. In the next section I shall focus on those attempts which try to answer this challenge.

3.4 Elaborating on the Own Mode of Presentation Claim

The accounts presented so far share the view that phenomenal concepts and their referents are distinct entities, and these referents are connected by a causal link to the application of phenomenal concepts. This, however, seems to pose a problem: if phenomenal concepts and their referents are distinct entities then they might occur independently of each other. It seems conceivable, then, that one deploys the phenomenal concept RED*[30] when having greenish experience (or without having any colour-experience at all).[31] This is highly counter-intuitive though: RED* is supposed to be the very vehicle of thought one deploys when one thinks about a reddish experience. That is, intuitively, anybody who introspectively deploys the

[30] Where RED* is the specific recognitional/demonstrative concept picking out the subjective phenomenal quality of reddish experiences. Note that RED* is different from the phenomenal use of the public language word 'red' (cf. §3.3.2)—RED* is supposed to be the special vehicle of thought one deploys when one thinks about one's reddish experience. It is a further question if one associates this phenomenal concept with information about red objects and hence with the public language word 'red' (as non-inverts would normally do), or with, say, information about green objects and hence with the public language word 'green' (as colour inverts would do). See more about phenomenal concepts and associated information in §3.5. Stoljar (2005) clarifies what kind of concepts one might have related to red stimuli. The concept RED* as used here corresponds to his RED SENSATION.

[31] Distinct entities linked only by causation are metaphysically independent, i.e. there are metaphysically possible worlds where the laws of nature are different, and thus the causal link in question doesn't exist (under a categoricalist understanding of laws of nature—cf. Footnote 27 in §1.4.1). Cf. Balog (2012a).

phenomenal concept RED* in order to pick out an actual experience is actually having a *reddish* experience (cf. Balog, 2012a).

This suggests that there is a tighter, more intimate link between phenomenal concepts and their referents than ‘brute’ causation (cf. Carruthers, 2004; Balog, 2012a). This tighter link is what Loar tried to capture by his OMP-claim: phenomenal qualities, when picked out by phenomenal concepts, provide their own modes of presentation. That is, when one is thinking about one’s own reddish experience this deployment of the phenomenal concept RED* somehow involves the reddish experience itself. In what follows I shall introduce two accounts of phenomenal concepts which try to make it intelligible how such an involvement might work.

3.4.1 The higher-order experience account

Carruthers (2000b, 2000a, 2001, 2004) subscribes to the view that phenomenal concepts are *pure recognitional* concepts with no contingent modes of presentation. It is Carruthers’ main concern to show how such recognitional concepts are possible in a physicalistically acceptable way.

Phenomenal concept as pure recognitional concepts satisfy two separate requirements. First, phenomenal concepts consist in the capacity to recognise the type of experience they are concepts of. By deploying the phenomenal concept RED* one is able to recognise and re-identify the reddish experience-type. Second, phenomenal concepts have no conceptual connection with physical-functional concepts. One cannot acquire the recognitional capacity solely by acquiring knowledge in terms of physical-functional concept—not even if one knows everything there is to know in terms of these concepts. That is, phenomenal concepts are irreducible to, i.e. conceptually independent of, non-phenomenal concepts.

However, it seems that phenomenal concepts have conceptual connections with other phenomenal concepts. A typical example of this conceptual connection is the very fact that ‘this experience’ is the paradigm formulation of a phenomenal concept—i.e. that one is able to refer to one’s particular experience by the phenomenal concept ‘this experience’. When seeing something red, and deploying a particular phenomenal concept (say, RED* to refer to the reddish experience) one can know that the state one recognises as reddish is an experiential state. That is, one recognises the state one is in as an experiential state; one knows that ‘this state’ is an ‘experience’ where both terms are pure recognitional (and phenomenal) concepts. The former is a more concrete one whereas the latter is a more abstract one.

Where we are left is this. When presented with something red we are able to recognise it as an instance of *red* and we are also able to recognise it as an instance of *colour*. During recognition we deploy purely recognitional concepts. In the very same way, we are also able to recognise our experience (the reddish quality) as an instance of *reddish experience* and bring it under the concept RED* or as an instance of *experience* and bring it under the concept of ‘experience’. The purely recognitional concepts deployed in this higher-order recognitional task (recognising not physical stimuli but experiences of those stimuli) are phenomenal concepts.[32]

According to Carruthers, “a concept is recognitional when it can be applied on the basis of perceptual or quasi-perceptual acquaintance with its instances. And a concept is purely recognitional when its possession-conditions make no appeal to anything other than such acquaintance.” (Carruthers, 2004, p. 320) That is, when one deploys the first order recognitional concept ‘red’ or ‘colour’ one is perceptually (or quasi-perceptually, i.e. in imaginative re-creation) *acquainted with* a red object, which is by being directly present is available to recognitional classification. However, an analogous interpretation of the higher-order recognition of experiences is problematic for those who would like to reject qualia, the non-physical properties, which are directly present to the mind when deploying phenomenal concepts, and are subjects of recognitional classification. Accounting for acquaintance in physically

[32] See §3.5 for more on how the more abstract concept ‘experience’ refers.

acceptable (non-qualia) terms is what is needed in order to explain the nature of phenomenal concepts.

Moreover, it seems that this acquaintance-relation is what underpins the intimate connection between phenomenal concepts and their referents. As Carruthers formulates this intimate relation:

> "What I recognize when I deploy a recognitional concept of experience is in some sense presented to me (albeit non-conceptually) *as* an experience. I do not merely find myself judging '*This* is K', as it were blindly, or for no reason. Rather, I think that I am aware of, and can inspect and reflect on the nature of, the event which evokes that recognitional judgement." (Carruthers, 2004, p. 322)

That is, in a certain way, experiences present themselves as experiences. This is exactly Loar's Own Mode of Presentation claim, i.e. that phenomenal qualities, when picked out by phenomenal concepts, provide their own modes of presentation. According to Carruthers, then, providing an account of the acquaintance-relation—and thus accounting for Loar's OMP claim—is the main challenge for physicalist approaches to phenomenal consciousness. And, Carruthers argues, causal accounts are not up to the task. Instead, he proposes a so-called dispositionalist higher-order thought account (Carruthers, 2000a, 2004).

The approach Carruthers proposes is one of the many variants of the higher-order representational theories of phenomenal consciousness. All these theories agree that for a state to be conscious it should be suitably related to higher-order representations of that very state. The dispositionalist higher-order thought account claims that those mental states are conscious which are available in a non-inferential way to a 'theory of mind' or 'mind-reading' system producing higher-order thought, i.e. those mental states, which are "available to cause higher-order thoughts about [their] occurrence and content" (cf. Carruthers, 2004, pp. 330-331). This availability to the mind-reading faculty results in that the states in question, in addition to their first-order content, will have second-order content as well.

Let's say that one is actually staring at a ripe tomato and having the experience of seeing red. The first-order content of one's relevant perceptual state here represents the state of the environment, i.e. it represents the tomato as *red*. If this state is available to the mind-reading faculty then it will also have second-order content. The second-order content of the state in question is an experience-representing one: it represents the very state as an *experience of red*. The second-order content of the state represents the state itself as an experience of red in virtue of the fact that the mind-reading faculty contains a concept (a phenomenal concept) which is apt for referring to that state.

Carruthers argues that the mind-reading faculty[33] is an evolutionary product (cf. Carruthers, 2000b). It is a belief-forming system operating on the first order content of perceptual states. It lets its possessor understand the subjective nature of experience, it allows one to think not just about the object represented by a perceptual state but also about how that perceptual state represents the object. As Carruthers puts it, the mind-reading faculty makes it possible to "grasp the is/seems distinction" (Carruthers, 2004, p. 332). Once such a faculty is at work, one becomes capable of making judgements about one's experiences themselves. For example, one becomes capable of judging one's experience as of the reddish experience type. One makes this judgement in virtue of recognising the perceptual state in question as presenting the object as red. That is, these judgements are second-order recognitional judgements. (Cf. Carruthers, 2004, pp. 332-333)

To recap: a perceptual state whose cause under normal conditions is a red object, if available to the mind-reading faculty, has both a first-order content *red* and a second-order content *experience of red*. Just as one is able to form pure recognitional

[33] Which, for those favouring the massive modularity thesis, might naturally occur as a module in higher cognition.

concepts to recognise and categorise the object via the first-order content as of the red object type, one is also able to form pure recognitional concepts to recognise and categorise one's experience via the second-order content as one of the 'experience of red' type. This latter set of concepts is the set of phenomenal concepts.

This approach accounts for the intimate connection between phenomenal concepts and their referents, and thus for the OMP claim, in the following way. If a perceptual state with the first-order content *red* is available to the mind-reading faculty, it will automatically possess the second-order content *experience of red*. That is, the second order-content is completely parasitic[34] on (but distinct from) the first-order content. The perceptual state which presents the object as red, i.e. the experience of red itself gains the second-order content of *'experience of red'* by being available to the mind-reading faculty. Via this second-order content the subject becomes capable of categorising or re-identifying the experience, i.e. forming a phenomenal concept. This is the sense, Carruthers argues, in which an experience (a phenomenal quality) provides its own mode of presentation when picked out by a phenomenal concept.

3.4.2 The constitutional account

As we have seen, Loar explains how phenomenal concepts refer directly (the DR-claim) by claiming that phenomenal concepts have the phenomenal qualities they pick out as their own modes of presentation (the OMP-claim). Papineau (2002) draws attention to the fact that this claim is ambiguous. On the one hand, one might understand it as expressing the simple thought that when a phenomenal quality is referred to by a phenomenal concept, there is no further property contingently related to the phenomenal quality mediating between the referent (the phenomenal quality) and the concept (the phenomenal concept). In this sense, then, the OMP-claim simply rephrases the DR-claim, it adds nothing to it.

[34] That is, on Carruther's view, exercising the perceptual concept that represents objects as, say, red is a precondition on exercising the phenomenal concept RED*. Cf. §3.5.

On the other hand, though, one might recall the classical understanding of modes-of-presentation—the Fregean picture, where the mind by being able to think of certain properties uses this ability to form a term to refer to entities possessing those properties (cf. Papineau, 2002, p. 104). On this reading, what the OMP-claim would suggest is this: "the mind somehow already has the power to think about some phenomenal property, [...] and then uses this ability to form a mode of presenting that property" (Papineau, 2002, p. 104). But this is clearly nonsense. If the mind already has the power to think about a phenomenal property why would the mind bother to construct some further mode of presentation to enable itself to think about that phenomenal property. Or to put it the other way around: the claim that in thinking about a phenomenal property the mind relies on its ability to think about phenomenal properties is circular—it presupposes phenomenal concepts in the explanations of phenomenal concepts. That is, Papineau argues, the Fregean understanding of the OMP-claim ought to be rejected and a DR-analogous interpretation should be preferred.

Still, there is more to say about the OMP-claim which goes beyond the DR-claim. According to the so-called Constitutional Account (Hill & McLaughlin, 1999; Papineau, 2002; Block, 2007; Balog, 2012a) phenomenal qualities, when picked out by phenomenal concepts, provide their own mode of presentation *because* phenomenal concepts are partly constituted by the phenomenal qualities they refer to. Every phenomenal concept token—which picks out a phenomenal quality type—is partly constituted by a token phenomenal quality. That is, the token state that realises a token concept is also a token of the referent. Phenomenal concepts are thus special because the vehicles involved when they are deployed are special in that they are also (partly) the vehicles of the very experiences the phenomenal concepts refer to.

In the case of non-phenomenal concepts it doesn't matter what particular neural configuration constitutes a particular token of the concept in question as long as the requisite causal/informational relations between the neural realiser and the referent of the concept hold. Contrary to this, the constitutional account argues, in the case of phenomenal concepts, “constitution matters for reference, both in terms of how the reference is determined, and in terms of how the concept cognitively ‘presents’ its reference” (Balog, 2009, p. 306).

The constitutional account easily explains the intimate connection between a phenomenal concept and its referent, and how the phenomenal quality serves as its own mode of presentation. If phenomenal qualities partly constitute phenomenal concepts, then by thinking about phenomenal qualities in terms of phenomenal concepts one automatically gets acquainted with phenomenal qualities themselves: by tokening a phenomenal concept one also tokens the experience itself which is referred to by the concept.

This, however, rises an important question. “How do phenomenal concepts come to refer to experiences that they themselves exemplify? How does the constitution relation determine or partly determine the reference of a phenomenal concept?” (Balog, 2009, p. 308) The constitutional account focuses on phenomenal concepts being constituted by experiences. But constitution in general does not fix reference. Neither referents of a concept typically constitute the concept itself (e.g. the concept DAY-OLD CHICK is not constituted by day-old chicks), nor does a concept typically pick out its constituents (e.g. though according to a naturalist reading neurons constitute the concept DAY-OLD CHICK it does not refer to neurons but rather to day-old chicks).[35]

Some proponents of the constitutional account (Papineau, 2002; Balog, 2012a) explain this special feature of phenomenal concepts in terms of an analogy with how

[35] In fact, there are two separate questions here. The first-order question asks what the referent of a concept is, whereas the second-order question asks how the concept gets to have that referent. With regard to phenomenal concepts, Balog and Papineau agree on the first order question (the referent is the phenomenal quality, which, in fact, is a physical property), but disagree about the second-order question: Balog emphasises that reference is determined by constitution (Balog, 2012a), whereas Papineau argues for a causal-teleosemantic account (Papineau, 2002, 2007). See §3.5 for more on Papineau’s account.

quotation works. The so-called quotational approach draws attention to the fact that the specific connection between constitution and reference (an item partly constitutes another item which refers to the first item) is a fundamental feature of linguistic quotation. A linguistic item partly constituted by a pair of quotation marks and partly by something between the quotation marks refers to whatever there is inside the quotation marks. As Balog puts it: "In a quotation expression, a token of the referent is literally a constituent of the expression that refers to a type which it exemplifies and that expression has its reference (at least partly) in virtue of the properties of its constituent." (Balog, 2009, p. 308)

In line with this, the quotational approach to the constitutional account suggests that phenomenal concepts are composed of a perceptual state and an operator acting on it —i.e. a an 'experience operator' and a 'perceptual filling'. The operator has the structure 'the experience: ---' where '---' stands for a blank slot filled by perceptual states. Phenomenal concepts, then, are formed by such experience operators prefixing perceptual experiences. These terms, in turn, pick out what fills the blank slot in the operator that is, what the experience operator operates on (cf. Papineau, 2002, p. 117). As Papineau puts it:

> "The referring term incorporates the things referred to, and thereby forms a compound which refers to that thing. Thus, ordinary quotation marks can be viewed as forming a frame, which, when filled by a word, yields a term for that word. Similarly, my phenomenal concepts involve a frame, which I have represented as 'the experience: ---'; and, when this frame is filled by an experience, the whole then refers to that experience." (Papineau, 2002, p. 117)

So, simply speaking, phenomenal concepts refer to the item which fills the frame 'the experience: ---', i.e. on which the experience operator is applied. But this simple picture can't be right. Consider one of the textbook cases of deploying phenomenal concepts: the imaginative re-creation of a certain experience. In this case, for example, in the case when one imagines what it is like to see the colour of a day-old chick, what one does is imagining the yellow colour of a day-old chick and then applying the experience operator on it. The problem is that what gets 'quoted' in this case, i.e. what fills the blank slot in the experience operator, is an imaginative experience, not an original perceptual one. This is a problem since usually, when one imagines how it looked like when one last saw the yellow colour of a day-old chick and thinks about this experience with the aid of the 'the experience: ---' operator, what one will focus on, what one will be thinking of is the original experience of seeing the yellow colour of a day-old chick. The product of the imaginative re-creation—as Papineau puts it—is only a "faint copy" (Papineau, 2002, p. 118) of the original one. However, when deploying a phenomenal concept, one is reflecting on the original experience not this faint copy.

Papineau's solution is that phenomenal concepts deployed in imaginative re-creation do not refer to the particular (imaginative) experience actually quoted but to "any experience that resembles it appropriately" (Papineau, 2002, p. 118). Consider, for example, the master chicken-sexer, who even at home cannot stop thinking about chicken-sexing, and imaginatively re-creates a 'seeing a day-old-chick' experience in her mind while sitting on her sofa. The phenomenal concepts exercised when she is musing about what it is like to see, say, the colour of a day-old-chick refers to all those experiences which 'appropriately' resemble the one actually quoted.[36] To motivate this claim, Papineau relies on an analogy with how demonstratives like 'that colour' refer not just to the actual shade the utterer is pointing at but to a whole range of similar shades resembling the one actually present.

Given this resemblance claim, note that the quotational model itself does not explain the semantic power of phenomenal concepts, i.e. does not provide an account of how phenomenal concepts refer. Rather, it is exactly the resemblance claim which tells us what the referents of certain phenomenal concepts are: those experiential states

[36] Papineau argues that this 'resemblance account' can be applied to cases where one thinks about an experience one is actually having as well. The actual experience this introspective use of the experience operator prefixes, Papineau claims, is a *vivid copy* (rather than a faint one) of the experience of seeing yellow, because perceptual categorisation amplifies the actual perceptual experience. (Papineau simply assumes that one can consciously see, say, something yellow without seeing it *as* yellow—cf. Papineau, 2002, p. 121, Footnote 11.)

which resemble the actual experience being quoted by (i.e. filling the blank slot in) the experience operator. Surely, this is not much of an explanation of how phenomenal concepts refer until an account of resemblance is provided. For this purpose, Papineau turns toward a causal-teleosemantic theory of representation. As he puts it:

> "All that is needed is that subjects be disposed to use these terms to respond to such resembling instances in a uniform way, and perhaps that these dispositions have an appropriate history. On the causal or teleosemantic account of representation that I am assuming, it will be facts of this kind that determine the semantic power of terms which invoke appropriate resemblance to exemplars, whether or not the users of the terms articulate any ideas of such resemblances. In particular, it will be facts of this kind that will enable phenomenal concepts to refer to experiences which resemble 'quoted' exemplars appropriately." (Papineau, 2002, p. 119)

That is, the resemblance claim only tells us what the referent of phenomenal concepts are—it does not tell us *why* phenomenal concepts so refer. For an answer to this latter question, one should turn to the causal-teleosemantic theory: phenomenal concepts refer to those experiences which resemble the one actually quoted *because* whichever of these 'resembling experiences' is instantiated it causes the application of the same phenomenal concept, or because tracking the appropriate set of resembling experiences is the very function of the particular phenomenal concept (cf. Papineau, 2002, p. 121).

To sum up, Papineau's quotational variant of the constitutional approach to phenomenal concepts relies on two distinct ideas to account for the crucial features of phenomenal concepts. The quotational model itself (i.e. that an experience operator gets applied on perceptual experiences) accounts for how experiences provide their own mode of presentation when picked out by phenomenal concepts, whereas a causal-teleosemantic theory of representation accounts for how phenomenal concepts refer.

3.5 Phenomenal Concepts and Perception

In his more recent writings Papineau moves away from the quotational account of phenomenal concepts. He still thinks that phenomenal concepts has a close relationship with perceptual states, but no longer thinks that the analogy with linguistic quotational marks correctly captures this relationship. In this section I introduce his new account of phenomenal concepts, with a particular focus on how phenomenal concepts fit into the process of perception.[37]

3.5.1 Perceptual concepts

Imagine that a true novice joins our master chicken sexer—it's not just that she (the novice) has never ever tried to determine the sex of a day-old-chick, but she has never ever seen a day-old-chick before. So when she stands next to the master chicken sexer and sees a day-old-chick for the first time, she might form the thought 'oh, I haven't seen anything like *that* before'. However, after this first encounter, she will become able to imaginatively re-create the day-old-chick in her mind, and recognise it upon new encounters. That is, the novice can form thoughts like 'I wonder what *that* is called' (recalling the first encounter from memory), or '*that*'s what I have seen today' (skimming through the great handbook of chicken sexing later the evening of the first encounter).

The vehicles of thought deployed in these cases are so-called *perceptual concepts*. Perceptual concepts are the vehicles of thought making it possible for subjects to think about perceptible entities—entities featuring in one's perceptual experiences. They are formed upon the first encounter with a novel entity, and can later be

[37] Papineau motivates his shift from his original account to the new account by citing a challenge attributed to Tim Crane and Scott Sturgeon. The challenge draws attention to the fact that on Papineau's original account any exercise of a phenomenal concept "will demand the presence of the experience itself or an imaginatively re-created exemplar thereof" (Papineau, 2007, p. 112). But the problem, then, is that it seems impossible for Mary to think the following thought truly: "*I am not now having that experience (nor re-creating it in my imagination)*" (Papineau, 2007, pp. 112-113). In thinking this thought Mary would use the phenomenal concept just acquired. But Mary wouldn't be able to think such a thought truly on Papineau's original account, since according to that the exercise of phenomenal concepts "did indeed depend on the presence of the of the experience or its imaginative re-creation" (Papineau, 2007, p. 113).

activated upon new encounters, or when imagining the entity without its actual presence.

Papineau draws attention to that though it might be tempting to think of such perceptual concepts as some kind of demonstratives (since their deployment is typically described by relying on the demonstrative term ‘that’—cf. above), nevertheless they are not demonstratives. As we have seen, demonstratives pick out different things in different contexts (cf. §3.3.2). As opposed to this, perceptual concepts pick out the same thing no matter what the actual context of their deployment is. Perceptual concepts *track* certain perceptible entities—it is not the case that they pick out one entity at one time of deployment and a totally different entity at another time of deployment. To understand how they do so, it is useful to first see what Papineau thinks about the way perceptual concepts fit into the workings of the perceptual system.

According to Papineau, the perceptual system generates certain *sensory templates* on the basis of the information collected and conveyed by the sensory organs.[38] These templates, the realisation of which might be thought of as certain neural activation patterns, can be stored, and re-activated upon new encounters and imaginative re-creation. Such sensory templates serve as the perceptual basis for categorising, recognising, and ultimately tracking certain entities. The cognitive system is able to accumulate information about the entities these sensory templates stand for, and attach this body of information to the sensory template. Thus when the sensory template gets activated all this attached information gets activated as well[39] thus enabling the subject to interact with the entity not just on the basis of the information the subject’s perceptual system is able to collect during the actual encounter, but also on the basis of all the information that has been encoded and attached to the template

[38] Papineau here follows Prinz (2002). Cf. §5.1.2 and §5.2.1 for more on the architecture of and the exact processes within the perceptual and cognitive systems.

[39] Stored sensory templates might be activated in the sense that they get ‘resonated’ by certain incoming stimuli (cf. Papineau, 2007, p. 115).

during previous encounters. For example, the novice chicken sexer might not have seen the eyes of the day-old-chick when she first saw it and was concentrating on its cloaca. Nevertheless, upon her next encounter with the chick when she has time to take a good look at the eyes as well, she is able to attach this information to the sensory template, which in turn, from that moment on, will carry information about both the eyes and the cloaca of the day-old-chick (cf. Papineau, 2007, pp. 114-115).

A sensory template together with this body of attached information is what constitute a perceptual concept. That is, perceptual concepts consist of these two parts: the sensory template and the attached body of information.[40] Consequently, perceptual concepts are able to carry information from one use to another. It is exactly this feature, the capability to carry information from one use to another, which clearly distinguish perceptual concepts from demonstratives—demonstratives never carry information about the entity picked out by a previous use (cf. Papineau, 2007, p. 115).[41]

Given the idea that a perceptual concept accumulates information about a particular entity and makes it available for further use (guiding interactions, etc.) we immediately have a grasp on how perceptual concepts refer. The referential power of perceptual concepts is determined by their function to accumulate information about certain entities: a perceptual concept refer to that entity which it accumulates information about. If the perceptual concept the novice chicken sexer forms upon her first encounter with a day-old-chick carries specific information about idiosyncratic features of the chick (e.g. the specific shape of its bill etc.) then the perceptual concept refers to the particular chick. If however, a perceptual concept carries information like ‘has yellow fur-like feather’ and nothing specific to a particular

[40] Compare this to the files—with a P-slot for perceptual templates and an A-slot for abstract information—associated to concepts within the Fodorian framework. See §5.2.1 for a detailed discussion.

[41] As Papineau puts it: “Information about an entity referred to by a demonstrative on one occasion will not in general apply to whatever entity happens to be the referent the next time the demonstrative is used. By contrast, perceptual concepts are suited to serve as repositories of information precisely because they refer to the same thing whenever they are exercised.” (Papineau, 2007, p. 115)

chick, then it refers to the day-old-chick type. The point, as Papineau formulates it, is that "different sorts of information are projectible across encounters with different types of entities" (Papineau, 2007, p. 116).

That is, the referent of a perceptual concept is determined by the kind of information the given perceptual concept carries. If the kind of information carried by a perceptual concept is most appropriate to the day-old-chick type, then that perceptual concept will refer to the day-old-chick type. It the information carried is most appropriate to the particular day-old-chick, then the corresponding perceptual concept will refer to the particular chick (cf. Papineau, 2007, p. 117).[42]

Papineau argues that perceptual concepts form structured hierarchies (cf. Papineau, 2007, p. 117).[43] So, for example, in the case above, the perceptual concept of a particular day-old-chick adds extra information to the body of information of the perceptual concept picking out the day-old-chick type. Seen from the other direction, perceptual concepts picking out more and more abstract entities (say, birds, animals) can be formed by abstracting away from the specific information carried by the perceptual concept referring to the day-old-chick type.[44]

[42] Subjects might not be particularly good at recognising certain entities—that is, recognitional abilities can misfire. For this reason, Papineau thinks that Loar is wrong in thinking of perceptual concepts as recognitional concepts. The novice chicken sexer might be unable to discriminate between the day-old-chicks she sees during her first and second encounters, nevertheless, this must not entail that she cannot think about a particular chick only about the chick type (cf. Papineau, 2007, p. 117).

[43] Within such a hierarchy, the activation of a specific perceptual concept automatically activates all of those more general perceptual concepts which 'cover' the referent of the specific one (cf. Papineau, 2007, p. 117).

[44] Papineau further argues that subjects are able to detach the body of information originally attached to a particular sensory template, and store it in a non-perceptual 'file' which in turn accumulates information about the same entity the original perceptual concept does, but allows for non-perceptual thoughts about that entity. Such non-perceptual thoughts might be like the one the novice chicken sexer forms when later in the evening of the day of her first encounter with day-old-chicks she starts turning the pages of the great handbook of chicken sexing and thinks 'I wonder if that cute little thing is in here'—*without* imaginatively re-creating the day-old-chick. Papineau argues that such non-perceptual thoughts are possible and calls the concepts playing a part in forming them (i.e. the non-perceptual 'files') *perceptually derived concepts* (cf. Papineau, 2007, pp. 118-119).

3.5.2 Phenomenal concepts

Papineau equates conscious perceptual experiences with the activation of a certain range of perceptual concepts.[45] He also assumes that the phenomenal character of conscious perceptual experiences is determined by the sensory templates involved in the perceptual concepts the activation of which constitute the conscious perceptual experiences in question. That is, on Papineau's account, if, for example, the master chicken sexer deploys the same sensory template when thinking about a particular day-old-chick, which is also deployed by the novice chicken sexer when she thinks about the day-old-chick type, then the phenomenal character (the 'what-it-is-likeness') of the two experiences they are having will be the same (cf. Papineau, 2007, pp. 117-118).[46]

Phenomenal concepts—the mantra of this chapter says—are our vehicles of thought when thinking about the phenomenal character of conscious experiences. Given that this phenomenology is tied to the stored sensory templates, it is a natural idea that we use these very sensory templates to think about the phenomenology itself. This is exactly the main idea behind Papineau's new account of phenomenal concepts. As he puts it:

> "I want now to suggest that we think of phenomenal concepts as simply a further deployment of the same sensory templates, but in this case being used to think about perceptual experiences themselves rather than about the objects of those experiences." (Papineau, 2007, p. 122)

That is, when the novice chicken sexer first sees a day-old-chick she forms a corresponding sensory template. With the aid of this template she can think several different thoughts. For example, she can think about the particular day-old-chick. In this case, she uses the template to accumulate information about the particular chick

[45] What Papineau has in mind here is that activations at early stages of visual processing, though might qualify as perceptual concepts as defined by him, most probably do not constitute conscious experiences. Compare this with my core hypothesis in §4.2.3.

[46] See also in §5.2.

by collecting information appropriate to the particular chick and attaching it to the template. Alternatively, she can think about the day-old-chick type. In this case, she uses the very same template to accumulate information about the day-old-chick type by attaching such information to the template which is appropriate to the type. Or, Papineau argues, she can think about *seeing* a day-old-chick. In this last case, she uses the template to accumulate information about the conscious experience itself, i.e. she attaches experience-specific information to the template. This use of the sensory template, then, constitutes a phenomenal concept. That is, phenomenal concepts consist of sensory templates plus experience-specific information attached to them. And just as a perceptual concept consisting of a sensory template and some, say, a particular day-old-chick specific information attached to it refers to that particular day-old-chick, a phenomenal concept consisting of the same sensory template plus some 'the experience of seeing a day-old-chick' specific information attached to it picks out the experience of seeing a day-old-chick as its referent.[47]

This is Papineau's new account of phenomenal concepts. Note that this new account shares an important feature with Papineau's original quotational account: phenomenal concepts *use* an experience in order to *mention* it (cf. Papineau, 2007, p. 123). According to the original quotational account phenomenal concepts consisted of an experience operator of the form 'that experience: ---' plus a particular perceptual experience filling the blank slot of the operator. Within this framework, a phenomenal concept refers to an experience by using it in the sense of involving it inside the experience operator.[48] According to Papineau's new account, phenomenal concepts consist of a sensory template and some information specific (most appropriate) to that experience type. Within this new framework, a phenomenal

[47] Papineau argues that since particular experiences, unlike particular spatiotemporal entities, do not persist over time they cannot re-occur, i.e. subjects cannot re-encounter with particular experiences. Hence, experiences-specific information attached to a sensory template is always experience-type-specific, and thus phenomenal concepts refer to experience types rather than token experiences. Consequently, in order to pick out a particular experience, one cannot rely solely on phenomenal concepts—one needs to rely on further descriptions like 'the particular experience I am having now' (cf. Papineau, 2007, p. 123).

[48] To be more precise: phenomenal concept within the quotational framework refer to experiences *resembling* the one being involved inside the experience operator—cf. §3.4.2.

concept refers to an experience type by using it in the sense that an instance of that experience type goes with the stored sensory template and thus becomes activated whenever the phenomenal concept in question is exercised. In this sense, thus, Papineau's new account provides a similar explanation of how experiences provide their own mode of presentation when picked out by phenomenal concepts: the experience-type is mentioned by using an instance of the experience-type itself.[49]

I would like to close this chapter by drawing attention to how Papineau's new account anchors phenomenal concepts in the process of perception. In doing so, it sets the stage for the main part of this dissertation: the discussion of certain features of perceptual representations (cf. Chapter 4), and a detailed analysis of how these features can determine the special characteristics of phenomenal concepts, or even account for the classical target phenomena of Phenomenal Concept Strategy without relying on specific conceptual features—and thus provide an *alternative* to phenomenal concepts based approaches (cf. Chapter 5).[50]

[49] Papineau answers the challenge originally motivating him to turn away from the quotational account (cf. Footnote 37 in the beginning of §3.5) by claiming that just as it is possible to derive non-perceptual files from perceptual concepts it is also possible to derive non-perceptual files from phenomenal concepts. Such files stores and accumulates information about certain experience-types, without also including a sensory template. Since, as we have seen, the phenomenology goes with the sensory template, such derived files have no phenomenology. Analogously to perceptually derived concepts, these files might be called *phenomenally derived concepts*. They make it possible for subjects to think thoughts like 'I am not now having that experience (nor re-creating it in my imagination)', i.e. to think about experiences in a non-phenomenal way, since exercising these phenomenally derived concepts does not activate the experiences they mention.

[50] In this chapter I concentrated on introducing the fundamental idea behind, and the difference between the main versions of Phenomenal Concept Strategy. For arguments against this strategy, see Chalmers (2007), Levine (2007), or White (2007). For arguments answering the challenges raised by these authors and thus defending Phenomenal Concept Strategy, see, for example, Block (2007), Papineau (2007), Levin (2008), Diaz-Leon (2008, 2010), or Balog (2012b).

Part 2

The Monadic Marker Account

Chapter 4:
Monadic Markers

4.1 Three Observations about Conscious Experience

Imagine a new-born joey. Wait, you haven't seen one yet? No worries, I can tell you how it looks. Imagine a foetus with forelegs only. Oh, you haven't seen a foetus either? All right, imagine then a bean-shaped creature approximately the size of a lima bean. Now imagine that it has tiny arms: there are two cylinder shaped outgrowths on each side of the bean (these are the upper arms) and there is a cone shaped formation (the forearm) connected to each cylinder with an elbow. And imagine that it is pinkish.

4.1.1 Structure in experiences

Can a description like the one above help in getting a grip on what it is like to see a new-born joey? It seems to be a safe bet to say that it can. After all, it is quite similar to those everyday descriptions we often give when we would like to help someone imagine something unseen—and we usually succeed. In order to see what the relevant factors are in these descriptions, consider two thought experiments.

First, imagine a marsupial-expert, Josephine, who knows everything there is to know about marsupials: how they mate, how young ones are born, how they make their way up through the fur of the mother into the pouch, and so on. However, imagine that Josephine has never ever seen a joey; she has spent all her life locked up in a room deprived of all joey sights. She has been carefully trained: no photographs, no pictures, not even a sketch of a joey has ever been presented to her. Nonetheless, she has been given as detailed as possible descriptions about the shape, colour and size of all different kinds of marsupials including new-born joeys. Now consider this question: can Josephine imagine what it is like to see a new-born joey on the basis of these descriptions alone, without actually seeing one?

Second, consider the more extreme case, where Josephine is deprived of all shape sights. Imagine, for example, that Josephine has been given glasses at her birth imitating the effect of a cataract: she might see some colours, but not definite shapes —all visual experiences she can ever have are blurry. Furthermore, imagine that she has even been deprived of all tactile shape stimuli as well, just in case the answer to Molyneux's question (Locke, 1690/1987; Fazekas & Zemplén, 2005) turned out to be affirmative.[1] Now imagine, that our shape-deprived Josephine is provided with all the descriptions the original marsupial-deprived Josephine has ever been given, and consider the question if shape-deprived Josephine can imagine what it is like to see a new-born joey.

Intuitively it seems that the two cases above are different. In the first case, it seems true that Josephine knows (i.e. is able to imagine) what it is like to see a new-born joey. After all, in the first case, Josephine is in a position analogous to the position fellows of the Royal Society of the early 1800s were in, who learned everything possible about marsupials solely from the written reports of the expeditions sent to discover the new land of Australia—and it seems reasonable to suppose that they could imagine how, say, kangaroos or joeys looked like well before the first sketches arrived.

Contrary to this, in the second case Josephine seems to lack the relevant knowledge: it seems that she would learn something new if her glasses were taken off and a joey was presented to her. Intuitively, she wouldn't be in a better position even if she was not only a marsupial-expert (thanks to the descriptions given to her), but omniscient with respect to all the physical facts—if she knew everything there is to know about the physical world in terms of descriptions (say, if the book of all physical knowledge, famous Mary is using, had been read out loud to Josephine).

[1] Though see Held et al. (2011) showing that the newly sighted fail to match seen with felt.

Shape-deprived Josephine has never ever had a shape experience, and thus is unfamiliar with shapes in general. Just as Mary, who lacks all relevant colour concepts[2], shape-deprived Josephine lacks all the relevant shape-related concepts. This is the main difference between shape-deprived Josephine and marsupial-deprived Josephine: the latter possesses a lot of shape-related concepts ready at hand to be deployed when interpreting the descriptions given to her. Contrary to this, shape-deprived Josephine has no grounds whatsoever to grasp the meaning of the descriptions informing her about the shape of a new-born joey.[3]

The descriptions in question are useful because they inform the uninformed about a certain shape by decomposing the original shape into parts (simpler constituent shapes) and determining the relative positions of these parts. Describing a new-born joey as a lima been with two cylinders and two cones attached to it at the right places is useful only for those who are familiar with how beans, cylinders and cones look like. Marsupial-deprived Josephine is such a person, shape-deprived Josephine is not.

In other words, seeing a joey is a *complex* experience. It is complex in the sense that the content of the experience itself (i.e. what one experiences) has constituent parts—parts that one is able to discern within one's complex experience, and can be contents of experiences on their own right. One can have independent, stand-alone experiences of such constituents without necessarily experiencing other parts of the original complex experience, or without the presence of anything else at all in the experience (other than the constituent in question). So, for example, seeing a joey is a complex experience: one is able to discern and recognise a tiny arm or the colour pink within one's experience of seeing a joey. Moreover, one can undergo the experience of seeing tiny arms without they necessarily being connected to a lima bean shaped creature (or to anything at all), and similarly seeing something pinkish

[2] The relevant concepts in question, of course, are phenomenal concepts, not ordinary language colour concepts. Cf. §3.1.

[3] See §4.1.2 for a detailed discussion of why this is so.

might just as well be experienced on its own by, say, staring at a pink wall from 20 centimetres as in a Ganzfeld-experience (Metzger, 1930).[4]

Note that these parts freely recombine. One can have an experience of seeing a sphere instead of the lima bean with exactly the same tiny arms attached to it, or one can undergo the experience of seeing beans, cylinders, cones attached to each other in the right (joey-ish) sort of way but painted in blue. This is exactly what happens in the first case. Marsupial-deprived Josephine has never ever seen a joey, but since she is familiar with experiences of seeing beans, cylinders, cones, and the colour pink, she is able to recombine these in a way she has never experienced them before. Thus, on the basis of the description informing her how to combine the constituents, she is able to learn what it is like to see a new-born joey.[5]

In the second case, though, shape-deprived Josephine is not able to imagine what it is like to see a new-born joey. The reason for this is that she is not familiar with the relevant constituent parts. Since she wears cataract-glasses she has never ever seen beans, cylinders or cones. Nor has she seen anything more basic like straight lines, curves, etc. She might have ideas about what pinkish looks like but she is unable to combine seeing something pinkish with anything with a definite shape, no matter how detailed descriptions about the particular shape is given to her, simply because she has no access to experiences of simple shapes and thus she is unable to combine them into more complex shapes.

[4] For the sake of brevity, in what follows I will talk about e.g. the experience of seeing a tiny arm as a constituent part of the experience of seeing a joey, or, in general, about experiences being constituted by less complex experiences. By using this expression, however, I do not intend to imply that experiences are indeed constituted by other experiences. This way of talking should be understood as a shorthand for the case where one can discern and attend to a part of what one is experiencing—a part, which could be the content of a standalone experience on its own right.

[5] Note that the first person who has ever sketched an imaginary creature, say, a dragon or a unicorn was quite similar to marsupial-deprived Josephine (or to Royal Society fellows of the 1800s) in terms of being deprived of some complex (in this case, dragon or unicorn) sights. Still, she could imagine what it was like to see the creature in question, and produce the corresponding sketch.

Experiences have *structure*: there are complex experiences constituted by less complex parts. These less complex parts can have further constituents, which themselves might just as well be built up from still less complex parts, etc. For example, the experience of seeing a new-born joey has as its constituent part the experience of seeing a tiny arm, which in turn has as its constituent part the experience of seeing a cone-shaped formation.

There seem to be, however, a bottom level. At that level units constituting higher level complexes themselves have no further constituent parts—they are *simple*: none of their parts can be the content of standalone experiences. For example, the experience of seeing a new-born joey also has as its constituent part the experience of seeing something pinkish. Even if some (cf. Boynton, 1997; Tye, 2000) argue that seeing something pink is a complex experience having constituents like seeing something red and something white, it seems reasonable to argue that at least unique hue experiences—seeing something red, yellow, green, or blue (as seen in a Ganzfeld, for example)—are simple colour experiences.[6] They have no constituent parts whatsoever. Though unique hues themselves have certain fundamental features like hue, saturation and lightness (or brightness) these are not apt for being contents of stand-alone experiences. One cannot have an experience of seeing a particular hue without necessarily seeing saturation and lightness at the same time.[7]

That is, from the cases of the two (the marsupial-deprived and the shape-deprived) Josephines we can conclude on the following observations:

(O1) **Most experiences are structured.**

[6] See, for example, Hardin (1988) and Thompson (2000) for an argument claiming that all colour experiences are simple in the above sense.

[7] Cf. Jakab (2000) for a similar claim. Jakab tries to account for the ineffability of experiences in terms of absence of constituent structure of certain experiences. Note, however, that Jakab does not differentiate between claims about the structure of experiences and claims about the structure of representational states—he uses constituent structure in the sense of syntactic structure and ties it to representational atomism. Contrary to his account, here I treat phenomenological structure and representational structure separately, and it is a particular aim of my approach to establish a relation between the two. Cf. §4.2 and especially §4.4.1.

Typically, experiences are complex—they have discernible constituent structure with discernible parts that, in themselves, could occur as contents of standalone experiences.

(O2) **Some experiences are unstructured.**
Some experiences are simple, without any constituent structure—they have no discernible parts that could be contents of standalone experiences.

4.1.2 Structure and phenomenal character

So far we have seen that Josephine is able to imagine what complex experiences might be like solely on the basis of descriptions specifying constituent parts and their relative positions—if, and only if, she is already familiar with (i.e. has previously experienced) the constituents in question.

Note that what happens here is Josephine concluding on the phenomenal character of complex experiences on the basis of the phenomenal character of simple experiences plus descriptions. Complex experiences, just as simple ones, have phenomenal character. There is something it is like to see something red, and similarly, there is something it is like to see a new-born joey. What Josephine's example shows is that the phenomenal character of complex experiences can be accounted for in terms of the phenomenal character of experiencing their constituent parts[8] plus descriptions about the structure the constituent parts are organised into when forming the content of the complex experience.

These structural descriptions rely on spatial terms expressing the relative positions of the constituents. These spatial terms are the regular spatial terms utilised in forming physical descriptions, e.g. 'below', 'next to', etc. That is, it is possible to conclude on the phenomenal character of complex experiences solely on the basis of the phenomenal character of simple experience plus *physical* description.

[8] In the sense emphasised in Footnote 4 above in §4.1.1.

In fact, this is a quite interesting feature of having complex experiences. It seems that physical descriptions can convey relevant information about the discernible structure of conscious experiences. In other words, there seems to be some sort of overlap between phenomenal structure and physical structure. The experience of seeing a new-born joey is a complex experience with some internal phenomenal structure. As we have seen, the phenomenal character of such an experience is partly determined by a description (conveying information about the relative position of the constituents), which is formulated in purely physical (spatial) terms. To put it in another way: the phenomenal structure one can discern in one's complex experience (representing spatial structure) can only be characterised with the very same spatial vocabulary one deploys when one characterises regular spatial relations in physical descriptions.

To support this claim, consider the intuitive difference between the cases of Mary and shaped-deprived Josephine. Whereas Mary is able to make sense of the information about the stimuli resulting in regular colour experiences conveyed by the descriptions given to her, it seems that the same task is problematic for shape-deprived Josephine. Though Mary cannot imagine red, green, etc., she is perfectly able to understand concepts like surface reflectance, electromagnetic waves and wavelengths, and thus has no problems with the descriptions given to here. Contrary to this, since what shape-deprived Josephine cannot imagine are exactly those features, which play an important part in the very descriptions informing her about the physical bases of shape vision—e.g. an object having straight or curved edges, etc.—shape-deprived Josephine lacks the relevant concepts, which could help her in making sense of the descriptions provided to her.

That is, a shape-deprived subject cannot interpret descriptions characterising shapes. Shape experiences are quite unique in that their phenomenal characteristics 'resemble' in some sense the physical characteristics of the objects they represent.

Shape experiences are described by the very same vocabulary one uses to describe the shape of the object stimulus. Compare this with other types of experiences. Just as colour-deprived Mary has no problems using a colour-term neglecting vocabulary to talk about the features of lightwaves, a smell-deprived subject could just as well understand descriptions characterising the shape of molecules[9], or a sound-deprived subject the features (e.g. frequency, intensity) of compressed air. The fact that the vocabularies describing the physical features of the stimuli giving rise to colours, smells, etc. are different from the vocabularies describing the phenomenology tied to these experiences—whereas in the case of shapes and spatial relations in general there is no such different vocabulary—informs us about that, in a certain respect, attending to shape-percepts can tell us more about actual shapes (the object stimulus leading to shape-percepts, i.e. the spatial distribution of matter) than attending to colour-percepts can about actual colours (the object stimulus leading to colour-percepts, i.e. surface reflectances)—cf. Kulvicki (2005). There is something more to colours (and smells etc.) than what is presented by experiences—and only colour (etc.) science can tell us what it is (e.g. surface reflectance). In contrast with this, shape perceptions seem to reveal what there is to be known about shapes (the edges, corners, curves they have); we do not need shape science to do this.[10]

To recap: whereas the structure of colour-, smell-, or sound (etc.) experiences is unlike physical structure[11], the structure of shape experiences is ‘similar’ to the physical (spatial) structure of the objects triggering the visual system. This is in fact

[9] Or their vibrations (cf. Turin, 2006).

[10] Cf. also Jakab (2003, 2006) who argues that shape experiences are revelatory in a sense in which colour experiences are not. Note that one might want to argue here that the fact that for characterising the phenomenal structure of shape experiences one needs to rely on the very same vocabulary which is also deployed for the characterisation of regular spatial relations is only a sign of the *transparency* of experience rather than a clue to the nature of the structure of experiences. This line of thought, however, will not work, since the transparency claim applies to colour experiences as well—it seems that our experiences present objects as being coloured. However, there *is* a separate vocabulary for characterising the object stimulus of colour experiences, whereas such a distinct vocabulary is not available for shapes. That is, the point is not that objects and experiences of them can be characterised with the same vocabulary, but rather that there is an alternative vocabulary available for the characterisation of the object stimuli of certain experiences (e.g. colours, smells, etc.) whereas there is no such alternative available for other experiences (e.g. shapes).

[11] I.e. physical structure of the objects represented.

Locke’s observation about the distinction between primary and secondary properties (Locke, 1690/1987): whereas our ideas of primary properties (e.g. shapes in particular and spatial relations in general) resemble primary properties themselves, our ideas of secondary properties (e.g. colours, smells, etc.) do not resemble secondary properties.[12]

The moral is that for characterising the internal phenomenal structure of complex experiences we use the very same expressions and terms, which are usually deployed when characterising physical spatiotemporal structure. This is why purely physical descriptions conveying information about physical structure can really help someone imagine having a complex experience.

This structural aspect is definitely part of the phenomenal quality of conscious experience. Seeing a particular shape (say, the shape of a new-born joey) could not be like what it is without certain information about structure being conveyed in the experience. That is, physical descriptions can and do penetrate the realm of phenomenal characters—a significant amount of information about the phenomenal character of complex experiences can be depicted by physical descriptions. Consequently, there is no further explanatory (or in general: epistemic—cf. §2.2) gap related to the phenomenal character of complex experiences *given that one is familiar with* the *simple experiences* constituting the complex one in question. Physical descriptions do their work, and inform those who are acquainted with the items mentioned by the descriptions what it is like to see a complex experience (say a new-born joey).

Still, this structural aspect does not exhaust the phenomenal character of complex experiences, since simple experiences have phenomenal character as well. This is our third observation regarding structure in experiences:

[12] This Lockean resemblance-claim is often debated. See §5.3.1 for an extensive argument in favour of a distinction between primary and secondary properties.

(O3) **PhenQual (complex) = PhenQual (simple) + Structure**

The phenomenal aspect of having a complex experience is jointly determined by

(a) the structure simple constituents are organised into, and

(b) the phenomenal aspect of these simple constituents.

Note, that the structural aspect can wholly be captured by descriptions deploying solely physical terminology; it is the phenomenal aspect of simple constituent experiences that seems to resist reductive explanation.[13] Thus the epistemic gap—even in the case of complex experiences—stems from a gap between our knowledge of the phenomenal character *of simple experiences* and our physical knowledge.

4.1.3 Further defending the observations made

So far I have made three observations and formulated three corresponding claims. (O1) says that most experiences are structured in the sense that they have discernible constituent structure with discernible parts which could occur as contents of standalone experiences. (O2) says that there are some experiences which are unstructured in the sense that they have no constituent structure, no discernible parts that could be contents of standalone experiences. (O3) says that the phenomenal character of complex experiences is jointly determined by the phenomenal character of their constituents plus the structure these constituents are organised into.

Despite the arguments provided in the previous section, one might find these claims unsupported. That is, one might feel tempted to resist accepting these observations. In this section I investigate how one might reject (O1), (O2), or (O3)—and defend them against possible objections. In the course of this endeavour, I shall further clarify how these observational claims might best be understood.

[13] Cf. Chapter 6 and Chapter 7.

Take (O1) first. One might want to argue against (O1) by pointing out that the argument claiming that something is structured because it is made up from certain parts mixed together in a certain way is, in fact, inconclusive: it very well might be the case that though certain parts together constitute a new entity, this new entity is an indecomposable holistic mixture of its parts; i.e. the whole itself is unstructured. For example, if one adds salt to a glass of water and stirs it, then the resulting salty water does not seem to be structured in any relevant sense. Note, however, that I try to support the claim that there are structured experiences by thought experiments (marsupial-deprived Josephine) and real-world examples (fellows of the British Society of the 1800s) purportedly showing that it is possible for one to imagine having a previously unexperienced complex experience on the basis of a description mentioning experiential *constituents* (which have already been experienced on their own or as parts of other experiences the subject has previously had) *plus their structure* (typically spatial arrangement). The very fact that as a result of such an imaginative exercise (i.e. trying to imagine having an experience of, say, seeing a previously unseen object) one becomes able to recognise the very object in question shows that the constituent parts and their structure are discernible within the result. If the result were a novel indecomposable holistic mixture then one would not be able to recognise it solely on the basis of a description mentioning only its constituents and the way they are mixed.

Also note that (O1) is a claim about the phenomenology of having an experience of an object. It is the phenomenal character of a complex experience which is such that one is able to discern constituent structure and parts in it. Here the opponent, however, might try to resist again. For couldn't it be the case that those experiences I dubbed complex are, in fact, phenomenologically holistic and unstructured, and the structure one is able to discern is only representational structure? Such experiences would represent complex objects, but the way it is like to have them would be unstructured. As a first part of my answer, note that experiences which are phenomenally unstructured but nevertheless stand for complex structured objects are possible. In §4.3 I discuss in great detail how phenomenally simple unstructured experiences can represent complex structured objects.[14] However, experiences of this latter kind are such that one is *unable* to discern structure in them—after all, this is what being phenomenally unstructured amounts to. If internal structure is phenomenally unavailable for the subject in her conscious experience then it is impossible for the subject to discern any kind of representational structure in her experience. This is the very idea behind the transparency thesis, i.e. the claim that being aware of the phenomenal qualities of a conscious experience amounts to being aware of the qualities of the object the experience represents (cf. e.g. Tye, 2000),[15] and one of the main motivations behind representational theories of consciousness (cf. Lycan, 2008).[16] That is, the opponent's two conditions, i.e. having a phenomenally holistic and unstructured experience, and nevertheless being able to discern representational structure within the experience contradict with each other, and thus cannot be the case all at once. I conclude that there is discernible structure in certain experiences, and this structure is structure within the phenomenal character of those experiences.[17]

Take (O2) next. Here the fundamental claim is that there are some experiences in the case of which the subject is not able to discern any kind of constituent structure. The paradigmatic examples used were homogeneous unique hue experiences. Note again, that this claim is also about the phenomenal character of experiences. The central observation behind the claim is that when one considers what it is like to have, say, a yellow experience, then one is unable to discern constituent parts within the yellow experience in the same way one can discern constituent parts within a new-born joey, or a day-old-chick experience. However, one might want to object here that we

[14] See also §5.1.

[15] See also Footnote 7 in §2.1.1.

[16] Note, however, that even if I hold (O1), it does not commit me to representationalism. See §4.3 and §5.3.1 for arguments which run straight against representationalism.

[17] Note that my claim about the complexity and structure of certain experiences is compatible both with the unity of consciousness (cf. Brook & Raymont, 2010), and with such structure being dynamic (cf. Metzinger, 1995).

cannot make such an observation. We never really experience unique hues in themselves—rather we experience certain objects as being yellow. And in these cases it is not so straightforward that one cannot discern some internal structure within the phenomenology of one's experience. To overcome this obstacle, relying on homogeneous unique hue experiences as subjects can have them in a Ganzfeld-setting was my strategy above. However, one could argue (as, in fact, e.g. Metzinger and Walde (2000) have argued) that homogeneous colour experiences in a Ganzfeld-setting fail to qualify as 'experiences on their own right', since they typically vanish within 3-6 minutes (cf. Metzinger & Walde, 2000, p. 355). This, however, doesn't affect my argument. On the one hand, experiencing homogenous unique hues in a Ganzfeld-setting for 3-6 minutes is an experience good enough. On the other hand, and more importantly, it has never been a part of the line of thought above that one has to experience the so called simple (i.e. unstructured) experiences in themselves. One is able to imagine what it is like to see a day-old-chick even if one has never ever seen a day old chick, and never ever had a homogeneous yellow experience (never participated in a Ganzfeld experiment). What matters is having previous experiences within which one could discern those constituent parts which are mentioned in the descriptions used in the imagination task. Simple experiences, as it has been already noted, can freely recombine.

That is, the main point of (O2) is that when one discerns certain constituent structure within one's complex experience, then it is often the case that there is a hierarchy within this structure: certain constituent parts have further discernible constituents etc. However, there seem to be certain experiences which can be discerned as constituents of other experiences, but they themselves have no further discernible constituents. Experiencing yellow is such an experience. These experiences might be experienced on their own, as for example in a Ganzfeld-setting (for 3-6 minutes), but this is not a crucial requirement. What is crucial, though, is that they have no constituents which could be discerned and independently experienced even for 1 second.

Finally, take (O3). This last claim seems to be the most controversial and contra-intuitive of all, mainly because the very idea behind it—that one is able to conclude on the phenomenal character of a complex (yet unexperienced) experience solely on the basis of knowing the phenomenal character of its constituents (i.e. on the basis of previously experiencing these constituents either in separation or as parts of other complexes) and a description detailing how these constituents are arranged—seems to be questionable, and has to face with some straightforward counterexamples. Consider for example a Hermann grid (Hermann, 1870), and someone who has never ever seen a Hermann grid before (but has nevertheless had normal visual experiences involving rectangles, squares, black and white). Now run the following thought experiment: imagine this Hermann-grid-deprived individual who is given a description telling her about a grid-map with white streets (rectangles) and black squares between them, and then asked to imagine what it is like to see a Hermann grid. Will this Hermann-grid-deprived individual be surprised when she first has the chance to see a real Hermann grid? Most probably, the opponent would argue, yes, since even if she could imagine white streets bordering black squares, she would definitely not expect to see grey parches in those street crossings she is not directly looking at. That is, those knowing what it is like to see white rectangles and black squares will not know what it is like to see a Hermann-grid (which consists of nothing more than white rectangles and black squares), even if they are told how these constituents are arranged. The phenomenal character of the whole seems to be more than the phenomenal character of the parts plus structure.[18]

Note, however, that this is a counterexample to a claim which has never been intended to be implied by (O3). (O3) is not the claim that the phenomenal character

[18] Other nice examples, involving only basic shapes, might be the Zöllner illusion or the Poggendorff illusion (cf. Zöllner, 1860). In the case of the Poggendorf illusion, for example, a straight black line is crossing behind a white rectangle with black borders at an angle of 45 degrees. Anyone, who is Poggendorf-illusion-deprived might be expected to be able to imagine (if familiar with black lines and white rectangles) how the Poggendorf illusion would look like—except, that most probably they would be wrong: when actually experiencing the Poggendorf illusion subjects see the two parts of the black line on the two sides of the white rectangle as being offset rather than aligned.

of experiencing a complex object is jointly determined by the phenomenal character of experiencing *parts of the object* plus *the structure the parts of the object* are arranged into when forming the object. Rather, it is the claim that the phenomenal character of experiencing certain constituents discernible *as constituents of a complex experience* plus structural descriptions detailing the discernible structure *of the complex experience* jointly determine the phenomenal character of the complex experience in question. That is, the description provided should not be about how the object stimulus is structured, but rather how the complex experience of seeing the object stimulus is structured. So the alternative thought experiment—the one (O3) does rely on—goes like this. Ask the Hermann-grid-deprived individual to imagine a grid-map with white streets and black squares between them and grey patches in the street crossings except the one she is directly looking at. After given this description (and given that she has already seen white, black, grey, squares and rectangles) will the subject be surprised when presented with a Hermann grid? (O3) claims that no, she won't be. There might very well be special rules characterising how the human visual system integrates information which result in illusory percepts,[19] this, however, has nothing to do with (O3). (O3) is a claim about phenomenal character and phenomenal structure.[20]

4.2 Three Claims about Perceptual Representations

Having made the observations above about structure in experiences, in what follows I shall turn towards the representational underpinnings of conscious experiences with the purpose of trying to find a similar pattern in the cognitive-representational realm. That is, in the following sections I shall concentrate on whether there is anything

[19] In the case of the Hermann-grid, see Baumgartner (1960) and Schiller and Carvey (2005) for explaining the mechanisms underlying the perceived illusion.

[20] A similar answer can be given to any example relying on Gestalt effects and the Law of Prägnanz.

cognitive- and neuroscience can inform us about that might be corresponded to the observations made at the phenomenal level.[21]

With regard to the representational level, one might feel tempted here to object that what I have said so far is quite reminiscent of the obsolete theory of structuralism (e.g. Titchner, 1929). According to structuralism, complex mental phenomena are generated via assembling fundamental building blocks. For example, the building blocks of the perceived shape of an object are simple sensations with determinate quality, intensity, duration, etc. Structuralists claimed that the mind worked in accordance with the slogan 'the whole is equal to the sum of its parts'.

However, structuralism has soon been questioned on methodological grounds.[22] Moreover, Gestalt psychology (cf. e.g. Wertheimer, 1958) straightforwardly denied the basic tenet of structuralism and argued that the fundamental operation of the mind is holistic (slogan-wise: the whole is more than the sum of its parts). This holistic operation is based on organisational principles (law of closure, law of similarity, law of proximity, etc.). In short, according to gestaltism, a complex perceived shape is not generated from assembling simple parts but is perceptually primary—a direct result of the way the perceptual system operates. Therefore, the opponent might want to conclude, the way I have characterised the phenomenal quality of complex experiences must be flawed.

True, structuralism is problematic (however, mainly on only methodological grounds), and my account of the phenomenal character of complex experiences seems to be in tension with Gestalt principles. This tension, however, as we have

[21] This strategy is standard in scientific endeavours seeking for relations between different realms or explanations of certain phenomena in terms of other phenomena. See §7 for a quite detailed discussion of this issue.

[22] Structuralism relied on analytic introspection, a method which was supposed to provide scientifically reliable data about the basic constituents of experiences via introspective descriptions. However, analytic introspection was found to be unreliable: different subjects gave different descriptions even under similar experimental conditions.

already seen it, is only superficial.[23] Moreover, this objection misses the whole point of the previous section. What has been said so far is not intended as a theory of how perception works. (Remember, the debate between structuralists and Gestalt psychologists is about how the process of perception goes.) The claim of §4.1 is *not* that the process of perception generates complex experiences via combining simple experiences into a structure depicted by the physical descriptions in question. Rather, the claim is only that such structure is present in one's conscious experience. What the underlying mechanisms generating conscious experience are is another matter.

In the rest of this chapter, I am going to turn towards this question. In what follows, I will assume that there are certain perceptual representations corresponding to conscious experiences[24]. First, in the remaining part of §4.2 I shall make and defend three claims about the characteristics of the representations in question,[25] and then refine them on the bases of theoretical considerations coming from, and empirical evidences provided by state-of-the-art cognitive neuroscience. §4.3.1 and §4.3.2 shall clarify the consequences of the representational features §4.2 sheds light on. Finally, §4.3.3 shall introduce how the three observations of §4.1 and the three claims of §4.2 together lead to an account of conscious experience, and how this account fits into contemporary theories of consciousness.

4.2.1 Simple and complex representations

First of all, note that the observations presented in §4.1 match up with classical accounts of vision and object recognition (Craik & Lockhart, 1972; Marr, 1982; Biederman, 1987) quite well.

[23] See §4.1.3 and especially Footnote 20 there.

[24] §4.3.3 clarifies in exactly what sense perceptual representations correspond to conscious experiences.

[25] Neither of these claims are committed to problematic structuralist principles. In fact, both are compatible with modern approaches to the process of perception relying on multiple channels (F. Campbell & Robson, 1968) and even Bayesian inferences (cf. §4.2.2 and §4.2.3).

According to Biederman (1987), for example, the perceptual mechanism responsible for object recognition proceeds as follows. First, the perceptual system generates representations of primitive components and provides information about how these components are arranged. Then characteristics of the components and their arrangement is matched to representations of objects in the memory (where a representation of an object is basically a structural description expressing the relations among the components). The primitive components (called geons, standing for geometrical ions) are hypothesised to be simple, typically symmetrical volumes lacking sharp concavities, such as blocks, cylinders, spheres, and wedges.

Biederman's theory clearly has close ties to the fundamental idea behind structuralism. However, on the one hand, it is free from the problematic methodological toolkit of classical structuralism, and, on the other, seems to be compatible with modern, e.g. connectionist, approaches to mental representations.

In order to show this, Hummel and Biederman (1992) introduced a neural network model for object recognition, which was able to generate a viewpoint-invariant structural description specifying the object's parts and the relations among them. They proposed a network consisting of seven layers, such that the same output unit responded to an object regardless of (i) where its image appeared in the visual field, (ii) the size of the image, and (iii) the orientation in depth from which the object was depicted. Hummel and Biederman claimed that cells at the first two layers answered to features of component geons of the represented object. As they formulated:

> "Specialized connections in the model's first two layers parse images into geons by synchronizing the oscillatory outputs of cells representing local image features (edges and vertices): Cells oscillate in phase if they represent features of the same geon and out of phase if they represent features of separate geons. These phase relations are preserved throughout the network and bind cells representing the attributes of geons and the relations among them. The bound attributes and relations constitute a simple viewpoint-invariant structural description of an object. The model's highest layers use this description as a basis for object recognition." (Hummel & Biederman, 1992, p. 485)

Still moreover, in a further study Hayworth and Biederman (2006) argued that there is empirical evidence supporting Biederman's original theory. They presented fMRI studies suggesting that there are certain brain regions serving as the neuronal basis of object-part representations corresponding to geons. In particular, they claimed that neural representation of shape in the lateral occipital complex (LOC) is an intermediate one, encoding the parts (i.e. geons) of objects, rather than local features, templates, or object concepts.

To summarise, according to Biederman, in the perceptual system there is a hierarchy of representations (cf. also Craik & Lockhart, 1972; Marr, 1982) consisting of simple and complex shape representations. Simple representations—i.e. the representations of cones, cylinders and the like (geons)—are the basic constituents of all complex shape representations. Of course, constitution has a special meaning here. The claim that complex representations are constituted by simple representations is *not* a claim about the *vehicles* of these representations. Rather, the claim is that simple representations stand for constituent parts of what complex representations stand for.

The perceptual system also encodes the structural information describing how those parts simple representations stand for are organised into the objects complex representations stand for. This structural information (along with information about the simple constituent parts), then, enters the centre executive system and results in object recognition.

That is, simple perceptual representations constitute more complex perceptual representations in the following sense. In our cognitive system there is a hierarchy of perceptual representations ranging from simple through less complex to more complex representations. Less complex representations stand for features, which are parts of the features represented by more complex representations. Moreover, further processes are able to extract structural information from this hierarchy of representations encoding the arrangement according to which the parts less complex representations stand for are organised into the wholes more complex representations stand for. The centre system utilises simple representations and structural information together when recognises and categorises a complex object.

On the basis of all this, the initial claims concerning the representational hierarchy underlying conscious experiences can be formulated as follows:

(C1) **Most perceptual representations are structured.**

In the hierarchy of perceptual representations most representations are complex —they have constituent structure in the sense that there are primitive representations standing for parts of the complex object a complex representation stands for.

(C2) **Some perceptual representations are unstructured.**

In the hierarchy of perceptual representations some representations are simple —they have no constituent parts in the sense that there are no more primitive representations standing for parts of what simple representations stand for.

(C3) **Structured reps = Unstructured reps + structural info.**

For further processes the particular way complex representations are structured is determined by the set of the simple representations constituting (in the above sense) the complex one, plus the structural information expressing the relation between the features simple representations stand for.

4.2.2 Representational hierarchy according to contemporary cognitive neuroscience

The three claims above—(C1), (C2) and (C3)—has been formulated on the basis of classical high-level theories of object recognition (e.g. Biederman, 1987).[26] Note, however, that even if this approach is supported by low-level accounts (Hummel & Biederman, 1992) and even by neurophysiological data (Hayworth & Biederman, 2006) much more is needed to be said about whether the claims are compatible with state-of-the-art theories. The problem is that nowadays the particular view regarding representations that dominates Biederman-style approaches is generally considered as outdated.

In contemporary cognitive neuroscience interpretations of the results of experimental paradigms designed to uncover the representations corresponding to conscious experience typically rely on the signal detection theory (e.g. Kouider & Dupoux, 2004; de Gardelle & Kouider, 2009; de Gardelle, et al., 2009),[27] and internal processes within representational hierarchies are hypothesised in the light of Bayesian inferences (Kersten & Yuille, 2003; Lee & Mumford, 2003; Kersten, et al., 2004; Mamassian, 2006; Yuille & Kersten, 2006). In general, Bayesian inferences determine the probability of a particular hypothesis given a certain observation—i.e. how probable it is that the hypothesis in question is true given that a certain observational evidence is detected—on the basis of the prior probability of the hypothesis, plus the compatibility of the evidence with the hypothesis (the likelihood of the evidence in the light of the hypothesis).

[26] Biederman's geon theory is utilised here only as a paradigmatic example of the general classical approach which thinks of object representations in terms of specific configurations of basic parts. See Wallis and Bülthoff (1999) for a review comparing this (as they call it, 'LEGO land') approach with view- and image-feature based approaches.

[27] Signal detection theory claims that in forced choice paradigms the discrimination of signal from noise depends on two factors: (1) the discriminability index, which is a measure of internal response in terms of the separation and spread of the internal response probability density functions of the signal and the noise, and (2) the criterion location along the internal response axis determining a threshold of reporting signals. Bayesian decision theory provides resources to optimise the setting of this criterion location based on prior knowledge (typically gained by learning) regarding the probability distributions of the signal and the noise.

The essence of Bayesian approaches to perception is that within a hierarchy of representational states low-level cues make bottom-up proposals, which are validated by higher-level models (cf. Yuille & Kersten, 2006, p. 302). Imagine, for example the process of interpreting a set of written words executed by the visual system. This stimulus is processed through a hierarchy of representations ranging from very low levels where a fine-grained distribution of incoming visual energy is represented, through higher levels representing oriented edges, still higher levels representing segments, letters, words, and finally the meaning of the words. On the level of oriented edges feature detectors extract orientation information, determine oriented edges, and provide bottom-up proposals of possible segments on the basis of prior knowledge regarding the probability distribution of segments. These segment models then are deployed in a feed-back loop to calculate how well the proposed segment models explain the edge-features detected. Similarly, at the next higher level segments are grouped in accordance with prior knowledge regarding the probability distribution of letters and letter models are proposed in the feed-forward channels to the next higher levels. These letter models then are used for determining how well they explain the image features, etc. until the processing reaches the highest level.

This picture of a hierarchical organisation of representations matches up nicely with the classical claim that there is a representational hierarchy ranging from simple to complex representations where simple representations are the primitive constituents of the complex ones in the sense that simple representations stand for parts of objects complex representations stand for.

However, the claims of §4.2.1 (especially C1 and C2) look problematic in the light of the Bayesian approach. The problem is that within a Bayesian framework representations at every level are unstructured for the further processes of the next higher level. The only information representations at the level of oriented edges carry for processes at the level of segments is that particular edges with certain orientation are present in the visual stimulus. Processes at the level of segments cannot extract any further information about the visual stimulus from the representations of oriented edges. Similarly, the only information processes at the level of letters can extract from lower level representations of segments is that certain segments are present in the stimulus. Errors regarding oriented edges are lost at this stage—they are screened off by representations of segments. The same happens at higher levels as well: at the level of words lower level letter representations are interpreted as mere markers of certain letters being present in the stimulus, and processes have no longer access to the errors regarding segments. That is, it seems that the distinction according to which there are lower level simple unstructured representations and higher level complex structured representations evaporates once one subscribes to the Bayesian framework.

4.2.3 Refining the claims: representational hierarchy and central access

The problem, any attempt seeking for a pattern similar to the one depicted by (O1)-(O3) within the realm of representations underlying conscious experience has to face with, then, is this. On the one hand, it seems that some kind of a distinction between unstructured and structured representations ought to be found within the representational realm, whereas, on the other hand, the most influential contemporary approach to perceptual representations—Bayesianism—seems to be incompatible with this distinction.

However, I believe that by introducing an additional hypothesis this discrepancy could be resolved. First, note that the prospects of my proposal hinge on the right interpretation of what has been called 'further processes' in §4.2.1. Remember, according to (C1) most perceptual representations are structured, whereas according to (C2) some are unstructured *for further processes*. The apparent incompatibility between the claims (C1)-(C3) and the Bayesian framework is due to the interpretation of these further processes as processes at the next higher level in the hierarchy of representations.

Now compare this interpretation with what we have already seen with regard to Biederman’s original theory, namely that it is the processing of the central executive system that results in object-recognition: according to Biederman’s theory, geon-characteristics and structural information describing their relative positions enter the central executive system where they are matched to stored viewpoint-invariant representations of objects. Similarly, the idea that the processes correlating with conscious experiences involve at least accessibility for central systems like working memory (Baddley & Hitch, 1974) or the global workspace (Baars, 1988) is prevailing in almost all contemporary cognitive and neurobiological theories of consciousness (Baars, 1988; Prinz, 2000; Dehaene & Naccache, 2001; Dehaene, et al., 2006; de Gardelle & Kouider, 2009; Kouider, 2009).[28]

That is, in the claims that some perceptual representations are structured whereas others are unstructured for further processes, this last term ‘further processes’, instead of being interpreting as processes of the next higher level within the hierarchy of perceptual representations, should rather be interpreted as processes of the central systems like working memory or the global workspace.

With this clarification in place, it is possible to derive a genuine distinction between structured and unstructured representations—if one adds the following further assumption. Suppose that within the hierarchy of perceptual representations there is a range within which all stages of representations are simultaneously and independently accessible by central processes. Although representations at each stage are unstructured inputs for the processes at the next stage, representations at earlier stages stand for parts of what representations at later stages stand for, and central processes are able to extract this structural information via their access to all the levels of these representations within said range. Therefore, *for central processes* representations at later stages of this range are structured.

[28] See however Lamme (2006) or O’Regan and Noë (2001) for alternative approaches.

However, in accordance with the assumption added, central processes do not have access to all levels of representations of the full hierarchy. Only those representations can be accessed by central processes, which fall within the range in question. Consequently, even if there are lower level representations standing for constituents of the features representations at the earliest stage of the range stand for, central processes have no access to them. Therefore, representations at the earliest stage of the range are *unstructured for central processes*—working memory or the global workspace cannot extract structural information about the features they stand for. These representations are the most basic meaningful units, which carry information for central processes; none of their properties can independently be interpreted by processes in the central system. They screen off for working memory or the global workspace all the features (errors regarding these features, i.e. all other possible interpretations that might have been resulted in different bottom-up proposals) represented at levels lower than the earliest stage of the accessible range. Those representations are unstructured, then, which are at the entry-level of that range of the hierarchy of perceptual representations, which is accessible for central processes.

Note that the underlying characterisation that perceptual representations form a hierarchy, as we have already seen it, is unproblematic even in a Bayesian framework: modern approaches hold the view that perceptual systems process stimuli through a hierarchy of representations from very low levels where incoming energy distributions and local features are represented, through intermediate levels with viewpoint-centric representations of segments of integrated features, and shapes, to high levels of viewpoint-independent representations of objects and their meaning.

That is, with the aid of a Core Hypothesis about how central processes access the hierarchy of perceptual representations it becomes possible to establish a distinction between unstructured and structured representations.[29] This Core Hypothesis says the following:

> **Core Hypothesis**
>
> Central processes (working memory, global workspace) have access only to a range of the levels of the full representational hierarchy involved in the processing of perceptual stimuli.[30]

On the basis of this Core Hypothesis, the three claims related to the representations underlying conscious experience can be refined in the following way:

(C1*) **Most perceptual representations are structured.**

Within the accessible range of the hierarchy of perceptual representations all non-entry level representations are structured for central processes—they have constituent structure in the sense that there are other (lower level) representations also accessible for central processes standing for parts of the complex object a structured representation stands for.

(C2*) **Some perceptual representations are unstructured.**

Within the accessible range of the hierarchy of perceptual representations all entry level representations are unstructured for central processes—they have no constituent parts in the sense that no representations are accessible for central processes standing for parts of the object an unstructured representation stands for.

[29] See §4.3.2 for more on how unstructured representations resulting from the Core Hypothesis are related to other types of unstructured representations one can find within the cognitive-perceptual architecture. This comparison shall reveal the real importance of the Core Hypothesis.

[30] Note that though this hypothesis has been introduced as a mere assumption, it is not without any support. For example, modularist approaches to cognitive architecture claim that early level perceptual processing is encapsulated and inaccessible for the central system (Fodor, 1983; Carruthers, 2006). Similarly, Jesse Prinz (2000, 2002, 2007a, 2007b) argues in length that levels at the earliest stage of the hierarchy of perceptual representations lack accessibility by central processes in the relevant sense (see also Dehaene & Naccache, 2001; Dehaene, et al., 2006; Kouider, et al., 2010). Ultimately, the question whether this Core Hypothesis is correct or not is an empirical issues. It is the duty of further empirically-driven investigations to deliver judgment on this matter.

(C3*) **Structured reps = Unstructured reps + structural info.**

> For central processes the particular way structured representations are structured is determined by the set of the unstructured representations constituting (in the above sense) the structured one, plus the structural information extractable by central processes expressing the relations between the features unstructured representations stand for.

4.3 Monadic Markers

So far I have made and defended the claim that there is a sense in which perceptual representations can be classified into the two categories of structured and unstructured representations. Why this distinction is so important becomes clear if one examines the specific features unstructured representations have.

4.3.1 Unstructured representations

To fully appreciate the specific features of unstructured representations, let's compare them to the characteristic features of structured representations. Given that central processes have access to the representations of the constituent parts of complex objects structured representations stand for, and thus are able to extract structural information describing how the constituents of these complex objects are arranged, structured representations carry information about the physical structure of the objects they stand for. That is, structured representations present their referents for central processes as having structure.

Contrary to this, unstructured representations are such that they are the most basic, primitive representational components available for central processes, i.e. they have no further constituents in the sense introduced above: there are no more primitive, lower level representations accessible for central processes standing for constituents which themselves are parts of the whole represented by an unstructured representation. That is, it is impossible for central processes to extract structural information about the objects unstructured representations stand for. Consequently, unstructured representations present their referents for central processes as unstructured.

Though unstructured representations might stand for quite complex states of affairs and their vehicles might also be quite complex, for central processes of the system they are embedded in they appear as unstructured, *monadic* wholes. That is, it is neither the object represented nor the vehicle of the representational state that is unstructured, but the very way this representation is interpreted by the processes operating on it. Unstructured representations are the most basic 'meaningful' units carrying information for the rest of the system—none of their parts or properties can independently be interpreted by further processes (including processes of the central system).

In fact, the states of affairs unstructured representations stand for can be arbitrarily complex. Nevertheless, unstructured representations will not represent this internal complexity of the object they stand for—they only indicate the presence of the represented object. This is the basic function of unstructured representations: to indicate or *mark* the presence of complex states of affairs in the outside world. In contrast with structured representations, which due to the fact that their parts and properties can individually be interpreted by central processes are able to carry structural information, unstructured representations convey nothing else about the represented object except its presence. They are internally generated signals which might have distinctive properties helping further processes in distinguishing them and organising them into similarity spaces, but these features *screen off* the features of the represented objects. If the properties of an unstructured representation mapped the properties of the represented object,[31] then the representation would be able to carry information about features of the represented object. Since, however, it is not

[31] Kulvicki (2004, 2005) calls this kind of mapping isomorphism. §5.3.1 elaborates on this issue.

the case, the actual way they indicate the represented object is detached from the properties of the object. The only information further processes have access to regarding the object is its presence. That is, all further processes have is the unstructured representation as if it was a tag the system could use to indicate the occurrence of a certain state of affairs.[32]

Take, for example, the smoke signal indicating whether a new pope has been elected at the papal conclave. After the Pope dies, the College of Cardinals is summoned for a series of ballots to elect a new pope. After the ballots the papers used for voting are burned. The colour of the smoke signals thus generated indicate the result of the ballots to the people assembling in St Peter's Square. Dark smoke indicates that the ballot was unsuccessful, whereas white smoke indicates that a new pope has been elected. The smoke signals, as used here, are unstructured representation in the sense discussed above. First, the object they stand for—the actual process of balloting—is quite complex with a lot of features (e.g. number of participating cardinals, distribution of votes, identity of the new Pope elected, etc.). These features, however, are screened off by the properties of the signal itself: people in St Peter's Square cannot extract information about, say, the distribution of votes solely on the basis of the smoke signal. All they have access to are the properties of the signal itself. Note that except its colour, no other feature of the signal corresponds to features of the balloting process. Therefore, people do not pay attention to these other features of the signal. The molecular constitution of the smoke, the shape of the smoke cloud, etc. are all unimportant for those waiting for the results. It is only the colour of the smoke people really care about, and thus this is the only feature of the signals on the basis of which people organise them into a 'similarity space' (with white smoke indicating a successful, dark smoke indicating an unsuccessful event). That is, even if the signal itself is quite complex (with special molecular constitution, shape, etc.), for further processes—in this case, for people in St Peter's Square—it is unstructured. The smoke is the signal, and its colour is its only meaningful feature carrying information

[32] See §5.1 for a detailed discussion of how unstructured representations and screening off can be characterised in an information-theoretical framework.

about the event indicated. This is the signal's only meaningful feature because whereas it corresponds to a particular state of affairs, no other features of the signal map any other features of the ongoing events. Consequently, the only information the signal carries is that the particular state of affairs occurred.

That is, unstructured representations are such that they provide an informational cut-off point for the systems operating on them. There might well be a lot of information to be captured and carried (since the object or event represented is complex), and even the signal, the representational vehicle itself might very well be able to carry this information (since the representational vehicle is also complex), still, no further information is carried. Subsequent processes cannot extract further information from unstructured representations about the features of the complex states of affairs the occurrence of which is indicated by them. From a certain point of view this is a virtue of unstructured representations: they are tools for avoiding informational overflow within the representational system. For the sake of efficiency, the amount of information captured and represented by any system must stay between manageable measures. Unstructured representations might be considered as serving this purpose.[33]

4.3.2 Unstructured representations due to sensitivity threshold and monadic markers

In the section above I have agued that unstructured representations provide an informational cut-off point for the systems operating on them. Note, however, that for every cognitive-representational system there is a natural informational cut-off point regardless whether my Core Hypothesis holds or not. Any cognitive system which is not completely detached from its environment must possess some input mechanisms registering the physical stimuli coming from the environment and transducing them into signals the cognitive system is able to operate on. Now if an

[33] See §5.1 for more examples of unstructured representations, and a detailed account of how they work.

input mechanism has a sensitivity threshold below which it is unable to discriminate details in the incoming physical stimuli, then, necessarily, the signals (i.e. representations) produced by such an input mechanism when operating at its most fine-grained resolution will present the registered states of affairs as unstructured. Since the input mechanism is unable to register the internal structure of those features of the incoming stimuli which fall below its sensitivity threshold, the representations produced by the input mechanism will only indicated that there are such-and-such states of affairs present in the stimuli without carrying further structural information about the represented objects.

So, for example, the human eye has a spatial sensitivity threshold—a finite angular resolution of 50 white-black stripe-cycles per degree. This means that the eye is insensitive to white-black line pairs thiner than 0.35mm viewed from 1 metre (Russ, 2007). That is, all the fine-grained structure of an object (e.g. its molecular shape), which is below this resolution will be lost: the visual input mechanism will not be able to encode information about this fine-grained structure, and henceforth no features of the signal will carry information about it. The input mechanism will register all those stimuli as the same, which have the same coarse-grained (i.e. sensitivity-level) structure regardless of their fine-grained (i.e. below-sensitivity) structure. Consequently, details about this below-sensitivity structure is inaccessible for the system. For the system, features of the object at the sensitivity-level are such that no further details about them can be recovered. The representations standing for these features (even if their vehicle is complex) are unstructured for further processes.

That is, if an input mechanism has a sensitivity threshold, then the representations generated by this input mechanism will be unstructured (even for central processes). Note that these unstructured representations are at the lowest level of the full representational hierarchy—they are right at the entry level; they are generated directly by the input mechanism. Central processes have no access to representations standing for parts of what the specific representations in question stand for simply because there are no such ‘simpler’ representations within the system. These unstructured representations stand for the finest details the input mechanism is sensitive to.

So from this perspective, there is a way in which one can sensibly differentiate between representations that are structured and representations that are unstructured for central processes—even without my Core Hypothesis. Representations that are at the lowest level of the representational hierarchy are unstructured for central processes due to the sensitivity threshold of the input mechanisms, whereas all higher-level representations within the hierarchy of perceptual representations can be seen as structured for central processes. However, this distinction would not work for our present purposes. Remember, my aim here is to find a pattern related to the hierarchy of perceptual representations similar to the one observed with regard to conscious experiences—i.e. to find representational structures that could be corresponded to experiential states. Given this aspiration of the whole endeavour, the problems is this: unstructured representations at the lowest level of the representational hierarchy are very bad candidates for being the underpinnings of conscious experience. For example, Jesse Prinz, concentrating on the visual modality, argues that V1, the neural substrate of low-level visual representations, cannot be the ‘locale’ of consciousness (Prinz, 2000, 2007a, 2007b). As Prinz summarises:

> “The destruction of V1, the main neural correlate of low-level vision, apparently results in the loss of visual experience. Still, there is mounting evidence that V1 cannot be the locale of consciousness (see Crick & Koch, 1995; and Koch & Braun, 1996, for reviews). For one thing, visual hallucination can occur for a period after V1 has been destroyed (Seguin, 1886). Similarly, some subjects who experience blindsight after V1 damage continue to have phenomenal experiences in the blind fields under certain conditions (Weiskrantz, 1997). As Crick and Koch emphasize, V1 also seems to lack information that is available to consciousness. First, our experience of colors can remain constant across dramatic changes in wavelengths (Land, 1964). Zeki (1983) has shown that such colour constancy is not registered in

> V1. Second, V1 does not seem responsive to illusory contours across gaps in a visual array (von der Heydt, et al., 1984). If V1 were the locale of consciousness, we would not experience the lines in a Kanizsa triangle [...]. The fact that consciousness is lost when V1 is destroyed is, therefore, better interpreted as evidence that V1 is a primary source of inputs to another region in which consciousness can rightfully be said to reside." (Prinz, 2000, pp. 245-246)

That is, Prinz emphasises that the content of conscious experiences does not match up well with the features represented at the lowest-levels of the visual representational hierarchy. This claim generalises to all the different modalities. The lowest-levels of perceptual representations typically stand for incoming energy distributions and local features which do not appear in consciousness, and are insensitive to some of the information which is available in conscious experiences.

Instead of low-levels, Prinz, following Jackendoff (1987) and Crick and Koch (1995), argues that the best candidates for being the locale of consciousness are the intermediate levels of the representational hierarchy.[34] At these levels representations stand for viewer-centred shapes, illusory contours, constant colours, motion, etc. (cf. Prinz, 2000, pp. 245-246). That is, any attempt which seeks for identifying a pattern similar to the one found in conscious experience within the hierarchy of perceptual representations, should better to be able to locate this pattern at the intermediate levels rather than at the lowest level of the representational hierarchy.

This conclusion draws attention to the real value of the Core Hypothesis introduced in §4.2.3: it shows how it could be possible to make the distinction between structured and unstructured representations within that range of the representational hierarchy which seems to be the best candidate for being corresponded to conscious experiences. If there is a finite range somewhere between the low- and the high-

[34] Prinz also argues that high-level representations are also bad candidates for being the locale of consciousness. At high-levels, representations stand for e.g. object-centred shapes, picking out an object independently of the actual point of view. However, conscious experiences are viewpoint dependent: the way it is like to see a certain shape from a given point of view is typically different from the way it is like to see the same shape from another point of view.

levels of the representational hierarchy which is unique in the sense that relevant central processes have access only to this range of the representational hierarchy, then for these central processes representations at the entry level of this range will be unstructured, whereas representations at higher levels of this range will be structured.[35]

That is, from the perspective of being good candidates for underpinning conscious experiences, there is a crucial difference between unstructured representations at the lowest-level of the representational hierarchy (which are unstructured due to a sensitivity threshold) and unstructured representations at the intermediate level (which are unstructured due to a lack of central access to lower levels). To emphasise this difference, and to distinguish unstructured representations at the intermediate level of the representational hierarchy from any other kind of unstructured representation, from now on I am going to use the term 'monadic markers' for unstructured representations at the intermediate level. As we have seen above, this name captures two essential features of unstructured representations. On the one hand, they are *monadic* for the processes operating on them, since these processes cannot extract any further information about the objects they stand for, whereas on the other hand, they are *markers*, since rather than mapping the characteristics of the represented objects, they only mark their presence. According to this terminology, then, monadic markers are those perceptual representations at the intermediate levels of the representational hierarchy, which, in accordance with the Core Hypothesis, are unstructured for central processes.

4.3.3 The Monadic Marker Account of conscious experience

The first part of this chapter (§4.1) concentrated on conscious experience, whereas the second part (§4.2) focused on perceptual representations. In this final section of

[35] Note that Prinz' theory of Attended Intermediate-level Representations also emphasises the importance of accessibility for central processes. The kinds of relevant central access in Prinz' theory is access by attentional mechanisms and the fact that intermediate level representations 'broadcast' to working memory (Prinz, 2000, 2007a, 2007b).

this chapter, I shall introduce an account of conscious experience, which, on the one hand, is able to fit the results emerging from these two separate parts together, and on the other hand, possesses a significant amount of explanatory power with regard to the special characteristics of phenomenal consciousness.

In §4.1 I have argued for three claims related to conscious experience. I have pointed out that conscious experience is structured. Complex experiences have constituent parts, which are discernible for at least trained subjects, and can be independently experienced. Constituent parts can have further constituents etc., however, there is a bottom level with simple experiences which themselves have no further discernible constituents that could be contents of stand-alone experiences. Furthermore, I have claimed that the phenomenal character of complex experiences is jointly determined by the phenomenal character of their simple constituents and the structural information describing how the constituents are organised to form the complex. I have also argued that the internal phenomenal structure present in complex experiences of spatial relations is itself spatial. That is, the structural information one is able to extract from one's own spatial experience can only be characterised by the very same spatial vocabulary one deploys when one characterises regular spatial relations in physical descriptions.

Next, in §4.2 I have argued for three claims related to those representations, which are involved in the processing of perceptual stimuli. I have claimed that the representations in question form a hierarchy, such that representations at each level stand for parts of the features representations at the next higher level stand for. Then, I have hypothesised that central processes (working memory, global workspace) have access only to a subset of the levels of this hierarchy forming a range, and have further argued that central processes via accessing multiple levels of representations within said range are able to extract structural information related to the arrangement of the parts constituting represented features, due to which representations at higher levels of this range are structured for central processes. However, since such structure cannot be extracted with regard to representations at the entry-level of this range, these latter representations are monadic markers—representations at the intermediate level of the representational hierarchy, which are unstructured for central processes.

The two stories above share some important characteristics. In particular, observations (O1), (O2) and (O3) correspond to claims (C1*), (C2*) and (C3*) respectively.

> **(O1)-(C1*)**
>
> Subjects are able to discern structure present in their complex conscious experiences; similarly, central processes are able to extract structural information from perceptual representations within the range central processes have access to.
>
> **(O2)-(C2*)**
>
> Subjects are unable to discern structure in their simple conscious experiences; similarly central processes are unable to extract structural information from monadic markers (representations at the entry-level of the range in question).
>
> **(O3)-(C3*)**
>
> The phenomenal character of complex experiences is determined by the phenomenal character of simple experiences plus the structure characterising the arrangement of the constituents (constituents of constituents, etc., ultimately simple parts) discernible by the subjects; similarly, the particular way higher-level representations are structured for central processes is determined by the lower-level representations (and further still lower-level representations, etc., ultimately monadic markers) encoding parts of what the higher-level representation encodes, plus the structural information extractable by central processes.

Note that the structural information central processes are able to extract from the accessible range of perceptual representations is information about the physical structure of the stimuli. Thus, on the one hand, representational structure encodes or carries information about physical structure. On the other hand, phenomenal structure is also closely tied to physical structure: as it has been argued for in §4.1.2, phenomenal structure—at least spatial structure as discernible in conscious experiences[36]—reflects, or resembles[37] physical structure. That is, via their link to physical structure, phenomenal structure featuring in (O1)-(O3) and representational structure featuring in (C1*)-(C3*) can be corresponded to each other.

The account of consciousness I am proposing here is based on these similarities. It recognises that the role simple experiences play in the phenomenal domain is the very same role monadic markers play in the cognitive domain. On the basis of this, the proposal claims that the phenomenal character of simple conscious experiences as defined in §4.1 correspond to how monadic markers, as defined in §4.2 and §4.3, are dealt with, or interpreted by central processes of the cognitive-representational system. To be more accurate, in line with how such correspondence-claims are typically formulated[38], the proposal takes the form of an identity-claim:

> **Monadic Marker Account of Conscious Experience**
> *The phenomenal character of simple conscious experiences is identical with how monadic markers are interpreted by central processes of a cognitive system.*

According to the Monadic Marker Account of Conscious Experience, those features of the stimulus are conscious, which are represented at levels of the hierarchy of

[36] Or more generally: structure related to primary qualities. See §5.3.1 for an extensive discussion of a monadic-marker-based account of the primary-secondary quality distinction.

[37] Again, §5.3.1 is going to shed more light on how this resemblance-claim might best be understood.

[38] See Chapter 7, and especially §7.2 for an argument to this effect.

perceptual representations falling into that particular range central processes have access to. Simple experiences correspond to monadic markers in the sense that the phenomenal character of simple experiences is identical with how monadic markers are interpreted by central processes. Similarly, complex experiences correspond to higher level representations, and their structure corresponds to the structure central processes extract from the range of representations they access in the sense that the phenomenal character of complex experience is jointly determined by the structure extracted, and by how monadic markers are interpreted by central processes.

Note that the Monadic Marker Account of Conscious Experience, as formulated above, is an identity claim connecting phenomenal consciousness with how certain perceptual representations (monadic markers) are integrated into a cognitive system. Identity claims of this kind (typically identities connecting consciousness with certain physical processes) are often called brute identities or strong necessities in the literature (cf. e.g. Chalmers, 2010b). They are claimed to be ad hoc, unsupported and unlike any other kind of standard identity claims deployed in scientific explanations. The Monadic Marker Account of Conscious Experience, as I shall argue in the following chapters, is, however, immune to these objections. In this chapter, I have shown, that it is not ad hoc and unsupported in the sense that it has been formulated on the basis of certain similarities between the phenomenal and representational domains. In Chapter 5 I shall argue that it is not ad hoc and unsupported in the sense that it can be put to the test: it can be used to account for such essential features of conscious experience as the fact that phenomenal qualities resist functionalisation and the occurrence of the epistemic gap. Furthermore, in Chapter 6 and Chapter 7 I shall point it out that it is far from being unlike any other kind of standard identity claim featuring in scientific explanations—on the contrary, the Monadic Marker Account as an identity claim plays exactly the same role standard identities play in scientific explanations, *properly understood*.

Before turning our attention to these issues, however, I would like to close this chapter with a final remark. The Monadic Marker Account in itself is not able to tell which perceptual representations are conscious rather than unconscious, i.e. it does not specify precise requirements perceptual representations need to fulfil in order to become conscious. For example, as we have seen in §4.3.2, the Monadic Marker Account says nothing about why unstructured representations at the lowest level of the representational hierarchy play no part in being relevant (in a direct way) to conscious experience—instead, it turns to an existing theory of consciousness and follows its guidance with regard to this question. That is, the Monadic Marker Account is not a full-blown theory of consciousness. Rather it is like a *universal plugin*: it needs a proper theory of consciousness, a 'host', which is able to designate a set of perceptual representations as a good candidate for being the 'locale' of conscious experience, and once such a set is given, the Monadic Marker Account can be plugged in into the particular host theory in question and help it to become more than a mere correlational claim, i.e. to produce real explanations—explanations of e.g. why subjects experience the particular phenomenal structure they do, why phenomenal qualities resist functionalisation, and why there is an epistemic gap. To see how the Monadic Marker Account is able to achieve all this, let's move on to the next chapter.

Chapter 5:
Monadic Markers in Action

5.1 Monadic Markers within Dretske's Framework

In the previous chapter the Monadic Maker Account has been introduced. This chapter focuses on the Monadic Maker Account in action. In what follows I shall investigate how it fits into different psycho-semantic frameworks, how it departs from Phenomenal Concept Strategy, and how it is able to be more than a mere correlational thesis—i.e. how it is apt for providing novel explanations of important features of conscious experience.

In 2005, Murat Aydede and Güven Gülzedere have published a lengthy but highly informative paper in Nous, which argues for a version of Phenomenal Concepts Strategy from a Dretskean information-theoretic viewpoint. Surprisingly, their account and the Monadic Makers Account, though stem from very different observations and are motivated by very different considerations, converge on quite similar claims. This makes Aydede and Güzeldere's account a natural starting point here.

5.1.1 The information-theoretic framework

Aydede and Güzeldere (2005) build on Fred Dretske's information-theoretic psycho-semantics (Dretske, 1981). According to the Dretskean framework, a source *s* by being *F* (i.e., some entity *s* having an attribute *F*) generates a certain amount of information which is proportional to the number of alternative possibilities eliminated (by being *F* instead of, say, *G* or *H*, etc.). A signal *r*—any event, condition, or state of affairs whose existence or occurrence depends on *s* being *F*—carries the information that *s* is *F* iff "the conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than 1)", where *k* stands for what the receiver of the signal already knows about the possibilities existing at the source (Dretske, 1981, p. 65). The signal carries the information that *s* is *F* in virtue of possessing certain properties—its so-called information carrying properties. So for example, the signal of an analog speedometer display is able to carry information about the speed of a vehicle, and does so in virtue of having certain properties, namely a needle pointing at a particular place on the dial.

Pieces of information can be nested in each other. As Dretske puts it, "[t]he information that *t* is *G* is nested in *s*'s being *F* = *s*'s being *F* carries the information that *t* is *G*" (Dretske, 1981, p. 71). For example, the information that a vehicle travels at 52 km/h nests the information that the vehicle's speed is between 51 km/h and 55 km/h, or the information that the vehicle in question is moving. Any signal (e.g. the speedometer) carrying the information that the vehicle travels at 52 km/h also carries these further pieces of information.

Certain pieces of information nest other pieces of information, but not all information is nested in further pieces of information. To capture this, i.e. the difference between two ways information can be encoded in a signal, Dretske introduces the distinction between digital and analog information. As Dretske formulates it:

> "A signal carries the information that *s* is F in *digital* form if and only if the signal carries no additional information about *s*, no information that is not already nested in *s*'s being *F.* If the signal *does* carry additional information about *s*, information that is *not* nested in *s*'s being *F*, then I shall say that the signal carries this information in analog form. When a signal carries the information that *s* is *F* in analog form, the signal always carries more specific, more determinate, information in both analog and digital form. The most specific piece of information it carries (about *s*) is the only piece of information it carries (about *s*) in digital form. All other information (about *s*) is coded in analog form." (Dretske,1981, p.137).

The speedometer of a vehicle, as we have seen, carries, for example, the information that the vehicle is moving. However, the speedometer does not present the vehicle as merely moving but as travelling at 52 km/h. Travelling at 52 km/h is a particular way of moving; it is a more specific piece of information. That is, the fact that the vehicle is moving is nested in a more specific piece of information, and therefore is encoded in the signal in analog form. The information that the vehicle travels at 52 km/h carried by the speedometer is not nested in any other piece of information—it is the most specific, most informative piece of information the speedometer is able to carry. That is, according to Dretske, the speedometer carries this information in digital form.

A signal is not apt for providing more detailed, more determinate information about those facts which it encodes in digital form. Imagine a digital speedometer which only displays round values ending in either 0 or 5. So, for example, when the vehicle's speed falls between 1 and 5 km/h the speedometer displays 5 km/h, when it falls between 6 and 10 km/h the speedometer displays 10 km/h, etc. This digital speedometer, just like the analog one, is able to carry the information that the vehicle is moving. Moreover, this digital speedometer, just like the analog one, carries the information that the vehicle is moving in analog form—nested in some more specific piece of information. Similarly, both speedometers are able to carry the information that the vehicle's speed is between 51 km/h and 55 km/h. However, whereas the analog speedometer carries this piece of information in analog form (nested in the more specific information that the vehicle is travelling at 52 km/h), the digital speedometer carries it in digital form. The digital speedometer cannot reveal any more specific information about the car's moving than that its speed falls between 51 and 55 km/h. This is the most specific, most determinate information it carries.

It is one thing that a signal cannot reveal more specific information than those pieces which it encodes in digital form. It is quite another matter whether the analog information nested in the digital information of a signal is extractable or not.[1] A mechanism extracts an analog information from a signal if it produces a new signal, which carries the analog information in question as the most specific information it

[1] Dretske himself does not distinguish between extractable and non-extractable analog information. See Kulvicki (2004) for introducing this distinction.

carries, i.e. encodes it in digital form.[2] That is, when a mechanism extracts an analog information from a signal, it digitalises this piece of information and excludes all other information carried by the original signal. In order to do so, the mechanism needs to be causally sensitive to the analog information carried by the original signal. We have seen that a signal carries an information in virtue of possessing a certain property. Therefore, the extracting mechanism needs to be causally sensitive to this information carrying property of the signal.

Imagine that I find the idea of the digital speedometer above very fancy, and would like to have a similar digital display in my car. I go to a car tuning service, where they build in a ‘speedometer-digitaliser’ into my car, on top of the analog speedometer. It is connected to the analog speedometer such that it detects the actual position of the needle. It is wired in a way that it displays 5 km/h if, and only if the needle is between the 0 and the 5 km/h marks on the dial (0 km/h mark excluded, 5 km/h mark included), 10 km/h if, and only if the needle is between the 5 and the 10 km/h marks, etc. Now when my car travels at 52 km/h the analog speedometer, as we have seen, carries the information that the vehicle’s speed is between 50 and 55 km/h nested in the more specific information that the vehicle travels with 52 km/h, i.e. encoded in analog form. My new gadget digitalises this analog information. The original analog speedometer is able to carry information about the speed of the vehicle in virtue of having a needle pointing at a particular place on the dial. The speedometer-digitaliser is sensitive to this property of the original signal. It extracts the information that the vehicle’s speed is between 50 and 55 km/h by registering that the needle is between the 50 and 55 km/h marks. The display of the speedometer-digitaliser produces a new signal (displays 55 km/h) which carries the extracted information in digital form—it is the most specific information the digitaliser display is able to carry since it is insensitive to the actual position of the needle. The digitaliser extracts this information by disregarding irrelevant features of

[2] Hence, information extraction is also called digitalisation.

the original signal, i.e. abstracts away from the actual position of the needle, and thus from the most specific information carried by the original signal.

However, not all analog information is extractable in the above sense. To see this, consider the following example. Imagine that a plane is landing at an airport, and a picture is taken when the plane is approaching the runway. This picture represents, or encodes quite a lot of information about the actual state of affairs: the exact shape of the plane visible from the point of view of the camera, the type of the plane (say, an Airbus A380), the airline logo on the fin, the colour of the painting on the plane's fuselage, its distance from the ground, etc. The most specific information (the exact shape of the plane with all the characteristics, etc.) is encoded in digital form. More abstract (i.e. less specific) information, e.g. that there is an airplane with four engines landing is encoded in analog form, since it is nested in the most specific information carried by the picture—the picture doesn't depict just an airplane with four engines, it represents an airplane with four engines, a characteristically placed flight deck, a very specific shape, etc. The picture communicates the fact that there is a plane landing by carrying some much more specific information.

Now imagine a buzzer system, which produces a buzzing sound with a specific pitch when and only when a situation exactly like the one represented by the picture occurs. Let's further assume that the buzzer system is insensitive to all those finer details, which are indistinguishable in the picture due to its finite resolution. Dretske argues that when the buzzer goes on and produces the sound with the specific pitch, it carries exactly the same information (digital as well as analog) as the picture (cf. Dretske, 1981, pp. 138-139). Still, there is an important difference between the picture and the buzzing as signals carrying information about the landing of the Airbus A380.

The picture represents the Airbus A380 as an airplane with four engines, a characteristically placed flight deck, a very specific shape, etc. It carries the information that there is an airplane with four engines in virtue of having certain properties (i.e. in the case of a printed picture, certain dots of paint forming characteristic patterns—P_{4E}). Similarly, it carries the information that there is an airplane with a characteristically placed flight deck in virtue of some other properties (other dots of paint forming different patterns—P_{FD}), etc. The picture carries that there is an Airbus A380 in virtue of having all these properties (P_{4E}+P_{FD}+...) at the same time. Any system which is sensitive to property P_{4E}, but not to P_{FD}, etc. is able to disregard irrelevant features of the picture, and extract (digitalise) the information that there is an airplane with four engines.

Contrary to this, the buzzing system represents the Airbus A380 as a buzzing sound. The buzzing sound has no other properties only its specific pitch. Thus the sound carries the information that there is an Airbus A380 in virtue of having the property P_{itch}. Since the information that there is an Airbus A380 nests the information that there is an airplane with four engines, every signal carrying information that there is an Airbus A380 also carries the information that there is an airplane with four engines. If one wants to build a system, which is able to extract this nested information, one needs to construct a digitaliser which is sensitive to the information that there is an airplane with four engines, but disregards the information that there is an Airbus A380. However, since P_{itch} is the only property of the signal, the digitaliser cannot be sensitive of anything other than P_{itch}. Now the problem is evident: the buzzing sound carries its most specific information in virtue of having P_{itch}, so anything sensitive to P_{itch} is sensitive to the digital information carried by the signal, and thus cannot extract a piece of analog information by disregarding all other pieces of information carried by the signal.

On the basis of all this, it is possible to formulate the definition of extractability. If a signal *r* having properties P_1, P_2,... carries in digital form the information that *s* is *F*, which nests the information that *s* is *G*, then:

> **Information-Extractability:**
>
> *r* carries the information that *s* is *G* in extractable form just in case there is some property of *r*, P_i, in virtue of which *r* carries the information that *s* is *G* but not the information that *s* is *F* or any other *Q* that nests the information that *s* is *G*. (cf. Kulvicki, 2004, p. 385)

5.1.2 Sensation, perception, and cognition within the information-theoretic framework

The information-theoretic framework works with the following picture of the sensory-cognitive architecture. The sensory system, which consists of transducers and pre-conscious, low-level processors connects the central cognitive system to the environment. Transducers convert the physical input (different forms of energy) into neural signals usable by the rest of the sensory system, which processes the information transformed by transducers. Sensory representations form the interface between the sensory system and the cognitive system—they are the outputs produced by the sensory system and the inputs consumed by the cognitive/conceptual system (which in turn controls behaviour). It is the particular sensory representation that transfers information about a distal layout (the object represented) for the conceptual system.[3]

These sensory representations carry a lot of details in digital form about the objects they represent, and, in addition, nested in the digital information, a lot of further information encoded in analog form. This information is made available for the central system, which extracts certain pieces of analog information, and encodes them in digital form. That is, from an information-theoretic perspective, the cognitive system is an analog-digital converter. As Dretske puts it:

[3] I follow Aydede and Güzeldere's terminology here. Aydede and Güzeldere's sensory system and sensory representations roughly correspond to the perceptual system and perceptual representations of Chapter 3 and Chapter 4. The representational hierarchy discussed in Chapter 4 is a hierarchy inside Aydede and Güzeldere's sensory system. The cognitive/conceptual system (Aydede and Güzeldere) roughly corresponds to the central system (Chapter 4). See Footnote 4 in this chapter for further clarification.

> “Perception is a process by means of which information is delivered within a richer matrix of information (hence in *analog* form) *to* the cognitive centers for their selective use. Seeing, hearing, and smelling are different ways we have of getting information about *s* to a digital-conversion unit whose function it is to extract pertinent information from the sensory representation for purposes of modifying output. It is the successful conversion of information into digital form that constitutes the essence of cognitive activity. [...] Cognitive activity is the *conceptual* mobilization of incoming information, and this conceptual treatment is fundamentally a matter of ignoring differences (as irrelevant to an underlying sameness).” (Dretske, 1981, p. 142)

That is, concepts, produced by the cognitive system, are representational states carrying the information extracted from sensory representations in digital form. If, for example, when an air traffic controller looks out of the control tower’s window and catches sight of an Airbus A380 landing, her sensory system makes a rich array of information available for her cognitive system representing fine details of the airplane (including the four engines, the place of the flight deck, the colour of the fuselage, the logo on the fin, etc.). Her cognitive system, in turn, digitalises pieces of the analog information available, forming concepts—and thus contributing to recognising and categorising the stimulus—like AIRPLANE, or AIRBUS.

Aydede and Güzeldere (2005) start off with this picture. They supplement it by distinguishing between three classes of concepts: *sensory concepts* like RED, *perceptual concepts* like ROUND, and *observational concepts* like AIRPLANE. The distinction between these three sets of concepts is based on the abstraction/ digitalisation distance between sensory representations and related concepts. As we have seen, concepts digitalise analog information nested in the most specific information carried by sensory representations. The more information is disregarded by a concept, the longer the abstraction/digitalisation distance is between the sensory representation and the concept.[4]

Observational concepts have the longest abstraction/digitalisation distance because these concepts deliver information that can be extracted from many sensory representations with widely different digital informational contents. The digital information carried by the concept AIRPLANE can just as well be extracted from visual sensory representations of two wings, a fuselage, a fin, etc., as from, for example, auditory representations of jet engine noise, airflow noise, etc.

Perceptual concepts have a shorter abstraction/digitalisation distance. Though they, just like observational concepts, do abstract away from details carried in the sensory representation, they are more closely connected to the particular sensory representation they are acquired from than observational concepts. As Aydede and Güzeldere puts it: "[t]he information necessary and sufficient for the correct application of these concepts [...] is normally contained in the sensory base from which they are directly acquired" (Aydede & Güzeldere, 2005, p. 208). That is, contra observational concepts, their necessary acquisition conditions tie perceptual concepts to specific sensory representations. Examples of perceptual concepts (in the visual case) are concepts of spatiotemporal relations, geometric figures, and shapes.

[4] Aydede and Güzeldere's sensory and perceptual concepts might be corresponded to Papineau's perceptual concepts (cf. §3.5). Note however, that whereas Papineau's perceptual concepts are at least partly constituted by a perceptual template (which might be corresponded to intermediate level perceptual representations as discussed by Prinz (2000); cf. 4.3.2), Aydede and Güzeldere's concepts are novel representational states extracting information from sensory representations. Aydede and Güzeldere's observational concepts further complicate the picture. As Aydede and Güzeldere argues (see below), observational concepts can be acquired via book learning. That is, they are definitely not perceptual concepts in Papineau's sense. Perceptual concepts in Papineau's sense are tied to viewpoint centric perceptual representations. Note that it is possible to draw a finer-grained conceptual hierarchy than that of Aydede and Güzeldere. In this hierarchy, ROUND$_{\text{ViewpointCentric}}$ might be seen as a perceptual concept in Papineau's sense: it answers to round objects (circles) from those vantage points where they do look round. ROUND$_{\text{ViewpointInvariant}}$ is not tied to a particular viewpoint: it answers to round objects from many different vantage points, even from those where they do not look round (cf. high level, viewpoint independent perceptual representations as discussed by Prinz (2000); see also §4.3.2). Finally, ROUND$_{\text{ModalityInvariant}}$ is not even tied to the visual modality: it is activated when one reads or hears about round objects. Observational concepts, as Aydede and Güzeldere characterise them seem to be similar to this last kind of concepts.

Sensory concepts have a minimal abstraction/digitalisation distance. Their digital informational content is part of the digital informational content of the sensory representations they are acquired from. In other words, sensory concepts disregard only a minimal amount of the digital information present in sensory representations —there is only a minimal loss of information in the process of digitalisation. The prime example Aydede and Güzeldere use for illustrating sensory concepts are colour concepts. When the air traffic controller catches sight of the landing Airbus A380 her sensory representation carries information about a vast amount of detail, including the logo—say, a white kangaroo on a red background. The air traffic controller's cognitive system automatically picks up the colour information and encodes them in the sensory concepts WHITE and RED. These sensory concepts digitalise almost all the information carried by the sensory representation about the colour of the fin. The small amount of information loss is due to the fact that whereas sensory representations always present a colour as a particular shade, sensory concepts disregard the shade-specific information.[5]

Aydede and Güzeldere claim that sensory concepts are special, and argue that by relying on the special features of sensory concepts it becomes possible to explain why the epistemic gap occurs. Therefore, their account is a version of the Phenomenal Concept Strategy. The argument they propose runs as follows.

Since the abstraction/digitalisation distance between sensory concepts and their sensory base is minimal—i.e. there is almost no loss of information in the acquisition of a sensory concept—the central cognitive system doesn't need to do much digitalisation. Almost no extra information is used and then discarded in the process of acquisition. As Aydede and Güzeldere puts it, the mechanism that mediates the informational link between the tokenings of a sensory concept and the instantiations

[5] Aydede and Güzeldere reserve the notion of concept for those cognitive structures, which are involved in diachronic recognitional and identificational tasks. Though we are able to discriminate between different shades of a colour synchronically, we cannot do this diachronically. This observation leads Aydede and Güzeldere to the claim that there is some information loss in the process of digitalisation between the sensory representation of a colour and the corresponding sensory concept (Aydede & Güzeldere, 2005, pp. 207-208).

of the property it refers to (the sustaining mechanism of the concept) is *not cognitive*. The acquisition and deployment of sensory concepts is innate and automatic, and more interestingly, brute and primitive: whenever a system occupies the relevant sensory states it is ‘triggered‘ to acquire the corresponding sensory concept. Sensory concepts are closely tied to their sensory base—so closely that they are necessarily acquired from the sensory representations of the properties sensory concepts denote (Aydede & Güzeldere, 2005, pp. 210-212).

Contrary to this, the acquisition of perceptual and observational concepts involve digitalisation. There is a certain amount of information (some in the case of perceptual concepts, and a lot in the case of observational concepts) used and then disregarded by the cognitive system. The abstraction/digitalisation distance of observational concepts is the longest; though they can be acquired from the sensory representations of the properties they denote, it is not necessarily so. Observational concepts can also be acquired by other means, i.e. book reading, or inference. Their sustaining mechanisms are cognitive.

In the Dretskean framework, the semantic content of a concept is that piece of information, which is *completely digitalised* by the concept (cf. Dretske, 1981, p. 185). A piece of information about an object is completely digitalised, if the signal, which carries it as its most specific information does so without this information being nested in some other piece of information about an intermediary structure.[6] In our present context, this is an important requirement, since sensory representations play an intermediary role between concepts and the properties denoted by the concepts. However, in the case of observational concepts, it poses no problem: observational concepts can be acquired in many different ways (via many different sensory representations), so it seems plausible that they track distal objects without

[6] “Structure *S* has the fact that *t* is *F* as its semantic content =

(a) S carries the information that *t* is *F* and

(b) S carries no other piece of information, *r* is *G*, which is such that the information that *t* is *F* is nested (nomically or analytically) in *r*’s being *G*.” (Dretske, 1981, p. 185) tracking hugely disjunctive proximal (intermediary) properties (cf. Aydede & Güzeldere, 2005, p. 214).

Not so, with sensory concepts. In the case of sensory concepts, the complete digitalisation of the distal objects fails. Sensory concepts are directly and necessarily acquired from specific sensory representations, thus they seem to carry the most specific information about their environmental source by carrying the most specific information about an intermediary structure, namely the corresponding sensory representation. This means, that the most specific informational content[7] of sensory concepts is dual—it consists of information about the external property denoted by the concept, and information about the properties of the intermediary sensory representation (cf. Aydede & Güzeldere, 2005, p. 216). That is, the information-theoretic framework seems to give the wrong result: if the semantic content of a concept is that piece of information which is completely digitalised, then the semantic content of a sensory concept is not the external property (what should be), but the corresponding sensory representation.

Aydede and Güzeldere solves this problem by arguing that the conceptual system needs to integrate information carried by sensory concepts with information carried by perceptual and observational concepts. Since perceptual and observational concepts have their semantic content anchored in the environment (i.e. they are deployed in the recognition and categorisation of *external* objects), there better be mechanisms in place within the cognitive system, which lock the semantic content of sensory concepts to the external part of their dual most specific informational content.[8] Again, when the air traffic controller catches sight of the Airbus A380, her cognitive system needs to integrate those features presented by the sensory concepts RED and WHITE, with the features presented by perceptual concepts like CURVED,

[7] The *informational content* of a concept is the information carried in digital form plus all the information nested in the digital content (Aydede & Güzeldere, 2005, p. 205).

[8] However, Aydede and Güzeldere do not specify the underlying mechanisms.

and observational concepts like KANGAROO, as the features of a distal object. As Aydede and Güzeldere put it:

> "It is the pressure exerted by our practical interests in having a *coherent* global representation of our external environment that forces the conceptual system to pick out the [external property] *redness* as the semantic content of RED." (Aydede & Güzeldere, 2005, p. 218)

As a next step, Aydede and Güzeldere argue that the main difference between perception of external properties and introspection of internal sensory states is the way they utilise the same structure.[9] As we have seen, the cognitive system picks up information generated at the level of sensory representations[10] automatically, and with a very little loss of information, thus forming certain representational structures within the cognitive system with a very short abstraction/digitalisation distance from sensory representations. These representational structures carry information about the particular internal sensory representations they are acquired from, *and*, as a consequence of this, about the specific external properties the sensory representations in question stand for. Thereby, they have a dual (most specific) information content.

Now when one perceives properties of external objects, one deploys these representational structures in such a way, which utilises the external element of the information content of these structures. This is the case, which has already been discussed—the deployment of sensory concepts.

Parallel with this, when one introspects one's internal states, one deploys the same representational structures in such a way, which utilises the internal element of the information content of these structures. Aydede and Güzeldere argue that this is the

[9] Note how similar this claim is to the one made by Papineau (2007)—cf. §3.5.2. However, the framework Papineau is using is more Fodorian than Dretskean. See §5.2 below.

[10] Aydede and Güzeldere argue that sensory representations are information generating sources on their own—they eliminate the possibilities of alternatives. When a concrete sensory representation is instantiated it eliminates the possibilities of other representations being instantiated (cf. Aydede & Güzeldere, 2005, pp. 223-225).

case of deploying *phenomenal concepts*. That is, on Aydede and Güzeldere's account, deploying a phenomenal concept amounts to utilising the same representational structure, which is utilised when one deploys a sensory concept, but the two mechanisms—i.e. the phenomenal (internal) and the sensory (external) utilisations—differ in which part of the information content of the representational structure they select as the semantic content of the concept in question.[11]

With all these at hand, it is easy to see how Aydede and Güzeldere's account fits into Phenomenal Concept Strategy. Phenomenal Concept Strategy claims that phenomenal concepts are special, i.e. unlike any other concepts, which ultimately results in their conceptual irreducibility.[12] Aydede and Güzeldere's account explains what is special about phenomenal concepts: they (together with sensory concepts, of course) have their semantics fixed by a direct and immediate informational link to sensory representations. Consequently, sensory/phenomenal concepts are independent of any other concepts: they cannot be defined in terms of physical and functional concepts, and no such concepts are involved in fixing their reference (since their sustaining mechanisms are non-cognitive). From this the conceptual irreducibility of phenomenal concepts follows; as Aydede and Güzeldere put it: "this means that sensory [and phenomenal] concepts cannot be derived from any other concepts or theories couched in them" (Aydede & Güzeldere, 2005, p. 231).

5.1.3 Monadic markers as the sensory bases of sensory/ phenomenal concepts

Within the cognitive architecture, Aydede and Güzeldere concentrate on the conceptual faculty. In their account, the emphasis is on the characteristics of sensory and phenomenal concepts—the explanatory work is claimed to be done by these concepts due to their special nature: namely that they track the environment via tracking internal structures (sensory representations). This, Aydede and Güzeldere

[11] Again, compare this with Papineau's view in §3.5.2.

[12] Cf. §3.1 for more details about conceptual irreducibility.

argue, is a unique feature—all other concepts[13] track their distal causes without tracking proximal sensory representations.

The consequence is that whereas sensory/phenomenal concepts digitalise the most specific information carried by their sensory bases, and thus have a minimal abstraction/digitalisation distance, observational (and perceptual) concepts digitalise only some aspects of the analog information nested in the most specific information carried by their sensory bases, and thus disregard a lot of otherwise available information. But what is the reason for this difference in the first place? Why do sensory/phenomenal concepts digitalise the most specific rather than only some less specific information carried by the relevant sensory representation?

In fact, Aydede and Güzeldere have an answer to this question. They claim that those sensory representations, which serve as the sensory base of sensory/ phenomenal concepts carry their analog information in a non-extractable form. We have seen in §5.3.1 that a signal carries its analog information content in a non-extractable form if the signal has no such information carrying property in virtue of which the signal carries that particular piece of analog information but not other pieces nesting the analog information in question. In other words, if the signal has no other property distinguishable for a digitaliser than the very property in virtue of which the signal carries the most specific information, then the digitaliser will not be able to digitalise the analog information carried by the signal. The best the digitaliser can do is to respond to the only property the signal has, and thus to digitalise the most specific information the signal carries. According to Aydede and Güzeldere, this is exactly the case with those sensory representations which form the bases of sensory/phenomenal concepts (e.g. the sensory representation of a certain shade of red). As they put it:

[13] In accordance with the classification Aydede and Güzeldere provides, these are observational and perceptual concepts.

> “[The relevant information carrying feature of the sensory representation] by carrying information about a surface’s being red$_{16}$, also carries the analog information that it has a spectral reflectance, or that it (just) reflects light at different wavelengths. These are nested in the information that the surface is red$_{16}$. But these pieces of analog information cannot be recovered or extracted from the signal, i.e., from whatever feature of the sensory representation carries the color information in question.” (Aydede & Güzeldere, 2005, p. 207)

The crucial point, as we have seen, is that the digitaliser is not able to selectively respond to a certain piece of analog information carried by the signal, because the signal has no information carrying property (which the digitaliser is sensitive to) other than the one in virtue of which it carries the most specific information. That is, sensory/phenomenal concepts cannot extract analog information from (and thus must digitalise the most specific information carried by) their sensory bases because the sensory representations in question have only one information carrying property the conceptual system is sensitive to—i.e. because they are *monadic* (unstructured) for the conceptual system. The sensory bases of sensory/phenomenal concepts thus are monadic markers.[14]

That is, in a certain respect, the Aydede and Güzeldere account is very close to the Monadic Marker Account: the two accounts agree that among our sensory/perceptual representations some are monadic, i.e. (in information-theoretical terms) carry analog information in a non-extractable form. However, the two accounts differ in many respects.

First of all, recognising that there are monadic markers amongst sensory/perceptual representations is, at best, only half of the story. Explaining what monadic markers are (cf. §4.3.1 and §4.3.2), and how it is possible that for central processes some perceptual representations are monadic, whereas others are structured (cf. the Core Hypothesis and §4.2.3) is a crucial part of the Monadic Marker Account. Aydede and Güzeldere do not really address this question. Or to be more precise, it is not clear

[14] Cf. §4.3.1 and §4.3.2.

whether they have anything more in mind than the ‘unstructured representations due to a sensitivity threshold’ argument (cf. §4.3.2). Formulated in terms of their vocabulary, the question is why it is the case that some sensory representations carry analog information in a non-extractable form, whereas others carry it in an extractable form. Aydede and Güzeldere’s answer starts in the following way:

> “One of the most basic truths about autonomous intentional systems is that they have to interact with their environment informationally. So they have to have information entry mechanisms. These mechanisms cannot deliver every piece of information in analog form, i.e., in a form that is always nested by some further more specific information. There will have to be a cut-off point about the most specific information the mechanism can provide about the environment.” (Aydede & Güzeldere, 2005, p. 211)

In this passage Aydede and Güzeldere argue that the digital information content of representations generated by the input mechanisms of a cognitive system is determined by the sensitivity threshold of the input mechanism. This is equivalent with the argument presented in §4.3.2 explaining how an informational cut-off point is established at the entry level of a representational hierarchy by the finite sensitivity of the corresponding input mechanism. However, Aydede and Güzeldere do not stop here. They continue by arguing that from this fact about entry level representations it follows that these representations must carry analog information in a non-extractable form.

> “If this piece of digital information carries the analog information nested in it in an extractable format, then there will have to be structural features of the output representation carrying the (total) digital information that nest this information. Then the same question arises about the digital content of these features and its format. This process cannot go indefinitely. At some point there will have to be representational features with digital informational content that nests the analog information carried by them in a non-extractable format, at which point the property digitally represented won’t be represented as having internal constituents.” (Aydede & Güzeldere, 2005, pp. 211-212)

Roughly, Aydede and Güzeldere argue along the following lines. If the analog information nested in the digital information content of (i.e. the most specific information carried by) the entry level representations is carried in an extractable form, then these representations must have features, which, taken individually, are responsible for carrying (in digital form) each piece of the extractable analog information but not the most specific information (cf. §5.1.1). Now the question Aydede and Güzeldere asks is whether the analog information nested in the digital content of the individual features is carried in an extractable form or not. By the very same reasons, there must be more fine grained features ensuring that the analog information is extractable, etc. To escape infinite regress, Aydede and Güzeldere conclude that there must be a point where the analog information is carried in a non-extractable form. That is, Aydede and Güzeldere argue that since no signal can be infinitely structured, there is always information which is carried by a signal in non-extractable form.

Note however, that there are two different issues here. The first issue is how specific the digital information content of the entry level representations is.[15] The second issue is whether the analog information carried by theses representations as nested in their digital content is extractable or not. Aydede and Güzeldere in the first quote of the previous page, and in line with the argument of §4.3.2, argue that the first issue is determined by the sensitivity threshold of input mechanisms. Then, in the second quote of the previous page, they argue that the second issue is determined by how detailed feature sets (i.e. information carrying property-structure) these representations can have. Note however, that the two issues are interrelated. Entry level representations simply do not carry information about those states of affairs, which are below the sensitivity threshold of the input mechanisms generating these representations. That is, the most specific information carried by these representations is finite, and thus a finite set of information carrying property-structure is sufficient for making all the analog information[16] in principle extractable.

[15] That is, how specific the most specific information entry level representations carry is.

[16] At least about the physical properties of the represented object, which is the important issue from our present perspective—cf. 5.3.1.

The very fact, that there is a cut-off point at the entry level, i.e. that the most detailed information an entry level representation carries is constrained by the sensitivity threshold of the input mechanism means that the input mechanism is insensitive to the fine-grained structure of the represented object. If so, then those features of an entry level representation in virtue of having which the representation is able to carry information about the finest-grained structure the input mechanism is sensitive to simply carry no further information about the below-threshold properties of the represented object. The finer grained structure (i.e. the structure below the sensitivity threshold) of the represented object can change arbitrarily, it will not affect the representation generated by the input mechanism (cf. §4.3.2).[17]

From all these observations regarding the account Aydede and Güzeldere put forth, two related conclusions follow. First, the answer provided by Aydede and Güzeldere to the question of why there are unstructured representations is no different from the 'due to a sensitivity threshold' argument introduced in §4.3.2. However, as we have seen it there, this argument is problematic, since the representational base it points at is a poor candidate for being the 'locale' of consciousness. Second, although the information-theoretical notion of 'representations carrying analog information in non-extractable form' might be corresponded to the notion of 'unstructured representations' as used in §4,[18] not all unstructured representations are monadic markers—i.e. good candidates for being related directly to conscious experiences.[19]

In other words, the first major difference between the Monadic Marker Account and Aydede and Güzeldere's approach is that the latter cannot provide a sufficient answer to the question of why some of those representations which are good candidates for

[17] That is, it is in principle possible that if a represented object has no physical properties below the sensitivity threshold of a relevant input mechanism, and the representations generated have sufficiently detailed information carrying property-structure, then these representations will carry all the information about the physical properties of the object in an extractable form. Cf. 5.3.1.

[18] Only roughly, though, since as we have just seen, when representations are unstructured due to a sensitivity threshold of the input mechanisms they might carry all the information they carry about physical properties in extractable form.

[19] Cf. 4.3.2.

being the ‘locale’ of consciousness are unstructured, whereas others are structured. The Monadic Marker Account contributes significantly to understanding this issue—from a certain point of view, solving this problem is one of the main motivation behind the Monadic Marker Account.

The second main difference is connected to a related issue. The Aydede-Güzeldere account presupposes not just that there are representations carrying analog information in a non-extractable way, but also that there are representations which carry analog information in an *extractable* way. As we have seen this latter point is equivalent with there being structured representations in our cognitive system. Aydede and Güzeldere do not argue for this claim at all, they simply accept that this is the case. However, as it has been pointed out in §4.2.2, it is far from trivial how one could support this claim in cognitive terms—especially given modern approaches to perception, like e.g. the Bayesian approach. As we have seen, the problem there is that within the Bayesian framework all representations are unstructured, i.e. none of them carry analog information in an extractable form.

That is, from the perspective of modern theories of perception grounding structured representations in cognitive processes is a challenge. The Aydede-Güzeldere account does not recognise this challenge—it leaves this question entirely open. Contrary to this, the Monadic Marker Account has a fundamental constituent, the Core Hypothesis, which explains how certain representations within an accessible range of the representational hierarchy (where representations are unstructured for the processes at the next higher level) could nevertheless appear as structured for central processes.

The Aydede and Güzeldere account suggests a picture, where the perceptual system provides an array of representations with rich digital information content nested in which they carry analog information in an extractable form. Some of these representations, however, are such that they do not have information carrying properties encoding individually all the analog information the representations carry, and thus their analog content is carried in a non-extractable form. Aydede and Güzeldere sometimes talk about the latter set of representations as they appeared at higher levels of abstraction, which suggests that representations carrying analog information in non-extractable form are based on representations carrying analog information in an extractable form.[20] This is in stark contrast with the picture suggested by modern (e.g. Bayesian) approaches. According to this picture, since in a sense every representation is unstructured (i.e. carries analog information in a non-extractable form), it is the representation of structure which must be based on monadic representations and their organisation.[21]

Interestingly, the Aydede and Güzeldere approach seems to be unable to account for this latter picture. In the Dretskean framework, representations carrying analog information about the structure of an object in an extractable form cannot be based on representations carrying analog information about the same structure in a non-extractable form. Since representations carrying analog information in non-extractable form have no information carrying properties encoding the structure in question, there is simply nothing further mechanisms could be sensitive to in order to decode said structure, and thus no further representations can be formed which could carry information about this structure in an extractable form. The Monadic Marker Account proposes a solution via its Core Hypothesis—the organisation of unstructured representations and the way they are accessed by central processes must be taken into consideration. That is, where Aydede and Güzeldere's theory fails the Monadic Marker Account succeeds: it is able to incorporate the picture of perceptual representations suggested by modern approaches.

Finally, the third major difference between the Monadic Marker Account and Aydede and Güzeldere's theory is that whereas the latter mainly concentrates on the

[20] Kulvicki (2004, 2005) argues for a similar picture.

[21] Cf. §5.3.1 and especially Footnote 38 there for more on this.

conceptual level, and provides a detailed account of how sensory and perceptual concepts work (thereby directly supporting the Phenomenal Concept Strategy), the former claims that the focus should be shifted to the features of perceptual representations themselves. On the one hand, this move brings the Monadic Marker Account closer to contemporary theories of cognitive architecture, and on the other hand, it makes it possible to account for features of consciousness independently of one's commitments regarding what concepts are and how they work. From the point of view of the Monadic Marker Account, the special conceptual features versions of the Phenomenal Concept Strategy (e.g. the Aydede and Güzeldere account) talk about are only symptoms—consequences of the fact that there are monadic markers among perceptual representations.

To show this, i.e. to support the claim that it is the Monadic Marker Account which points out the relevant level of analysis where the real explanatory work is done, the next section will demonstrate that monadic markers can do the job even in a psycho-semantic framework which assigns no special characteristics to phenomenal concepts.

5.2 Monadic Markers & Fodorian Psycho-Semantics

The framework which allows us to further investigate the capabilities of the Monadic Marker Account is Jerry Fodor's theory of concepts (Fodor, 1987, 1998, 2008). On Fodor's account sensory and phenomenal concepts are not in any way special. To see why this is so, first, let's briefly summarise the main characteristics of this approach.

5.2.1 General Fodorian framework

The framework this section works with is a generalised version of Fodor's account. According to this general framework, what one can find inside the conceptual faculty are concepts and files. Concepts are atomic Language of Thought symbols which are nomically locked to the properties they are concepts of. Concepts are associated with files which are placeholders for information about the property being represented by the concept. These files typically contain abstract information about the object/property the concept stands for, but perceptual prototypes also play an important role in concept acquisition.[22] Consequently, files can be characterised as having two 'slots', one for perceptual templates (P-slot) and one for abstract knowledge (A-slot). Typically, P-slots get filled by perceptual templates via perceptually experiencing the object in question. The content of an A-slot, on the other hand, is what one reports when explaining the meaning of a particular concept.[23] However, files are not constitutive of concepts—on the Fodorian view concepts are strictly atomic Language of Thought symbols. Nevertheless, concepts might have complex representational content (as most of them do) which might be accessible for further cognitive processing via the associated files if those are not empty.

Take the example of seeing a day-old chick. When one, upon seeing a day-old chick, recognises it and entertains the concept DAY-OLD CHICK what happens in one's conceptual faculty is the following. The concept DAY-OLD CHICK is an atomic Language of Thought symbol which is asymmetrically causally dependent on day-old chick occurrences. The file associated to the DAY-OLD CHICK concept has its P-slot filled with a perceptual template of a day-old chick, and its A-slot containing abstract knowledge of day-old chicks, like 'it is an animal', 'it has yellow feather' and so on.

Perceptual templates filling P-slots are typically, but not necessarily formed via perception—they might also be formed from abstract knowledge. Following Fodor and Pylyshyn, this latter case might be called *simulation of look* in imagery (Pylyshyn, 2002, 2003). According to Fodor and Pylyshyn, imagination is simulation based on concept deployment. Combining concepts with perceptual templates in the P-slot of their associated files are accompanied by combining these perceptual

[22] See especially in Fodor (1998 Ch. 6; 2008 Ch. 6) and also Margolis (1998).

[23] Note how extremely similar this is to Papineau's (2007) account—cf. §3.5.

templates. This is how one might acquire complex perceptual templates from abstract knowledge.

For instance, consider an ornithologist, who, as it happens, has not yet seen any day-old chick so far—not even pictures thereof. However, she knows everything that can be learned from descriptions about the look of day-old chicks. This knowledge, together with her previous encounters with other birds via which she could form perceptual templates of wings, claws, bills etc. make her able to imagine day-old chicks. What she does is entertaining the combined concepts YELLOW, FEATHER, LEG and so on. Given that the P-slots of these concepts have already been filled, entertaining these concepts is accompanied by visual memories of yellow, feather, leg, etc. Roughly speaking, the visual image she constructs on entertaining DAY-OLD CHICK is one that arises from remembering actual perceptions of feathers, legs, the colour yellow, and so on.[24]

Perceptual representations mediate between sensory input systems and the central cognitive system—they are outputs of quite complex mechanisms taking place within the sensory input systems, and inputs of the conceptual faculty. However, no matter how complex and structured the low level processes are, they remain hidden from higher cognition. According to the Fodorian framework, mechanisms within the input systems are *modular* and *encapsulated*. Though this claim in its strictest sense is almost certainly false, at least in a sufficiently modest form it seems to be compatible with both scientific evidence and our everyday experience.[25] In terms of everyday experiences, mechanisms within the input systems are modular and encapsulated to a degree which makes these processes inaccessible to conscious reflection. One might have detailed knowledge about what happens in one's visual

[24] Note two things. First, structural descriptions, i.e. descriptions instructing the ornithologist how to combine the concepts YELLOW, FEATHER, etc. are necessary for the ornithologist to succeed. Second, what is explained here within the Fodorian framework, is exactly what happens with marsupial-deprived Josephine in §4.1.1—she *simulates the look* of a new-born joey by combining her concepts SPHERE, CONE, PINK, etc. on the basis of a structural description.

[25] Cf. e.g. Kosslyn (1994), Bryson (2000) and McDermott (2001).

system when one has the experience of, say, seeing something red; nonetheless, one is unable to consciously reflect on actual processes of one's early visual system, nor can one consciously influence these processes in any way. For such an interaction to take place, a sufficiently direct and detailed information transfer would be required between these two levels of representation. This kind of information transfer certainly does not obtain between e.g. colour processing and propositional representations.

5.2.2 Causal role exchange and functional un-analysability

Now that the stage is set, let's see what happens if we supplement the above framework with the additional claim that some of the perceptual representations mediating between the sensory and the conceptual systems are monadic markers. To be able to evaluate this situation we first need to consider an important consequence of having monadic markers, i.e. unstructured perceptual representations in one's cognitive architecture.

This consequence is the possibility of *causal role exchange*: monadic markers, and unstructured representations in general, are able to exchange the causal role they play in a system. The very same unstructured representation might play different causal roles in the same system at different times, and also, different unstructured representations might play the very same causal role in the same system (or in different systems of the same type).

Intuition pumps back up this claim. In what follows two cases shall be considered: the case of colour versus shape experience inversion, and the case of role exchange between conceptual atoms versus sentences. First, consider the contrast between colour experience inversion scenarios and visual shape experience inversion scenarios.[26] It is quite easy to conceive of a colour experience inversion scenario

[26] Recall that in §4.1.1 we have seen that colours are good candidates for unstructured experiences, whereas shapes are paradigmatic forms of structured experiences.

where the colour of ripe tomatoes look to colour-inverted subjects the way the colour of grass looks to the rest of the population without there being any further difference in their cognitive architecture. Contrary to this, visual shape experience inversion scenarios are harder to conceive of. If, due to some optical distortion, a subject misperceives circles as squares this tends to give rise to mistaken inferences in her mind about the shapes of certain objects. A fairly complex change in subsequent processing is needed to straighten out all geometrical inferences related to squares and circles in the subjects mind; and this is only a very simple case of shape inversion.

Second, take two atomic symbols of Mentalese, X and Y. Suppose that, in subject A's mind, X is locked to spoons, and has an associated file containing a perceptual template of spoons and relevant abstract knowledge. Also in subject A's mind, Y is locked to knives and has an associated file with proper contents. However, in subject B's mind, X is locked to knives and is associated with the knife-file, whereas Y is locked to spoons and has the spoon file associated with it. If the Fodorian view of concepts (in which atomic concepts are locked to the properties they represent and are associated with relevant knowledge) is right, then the role switch just sketched seems straightforwardly possible. Contrary to this, the semantic inversion of sentences (complex linguistic representations) is an utterly strange idea. Imagine that the sentence “Budapest is the capital of Hungary” expresses a geographical fact for subject A, but for subject B it expresses the very zoological fact “giraffes are taller than dogs” expresses for subject A.

The moral that follows is that complex representations are much more tightly embedded in a system than unstructured representations. Two complex representations cannot assume each others causal role, for instance, without a significant reorganisation of the whole system, whereas unstructured representations seem to be able to exchange their causal roles freely.

Although causal role exchange between unstructured representations seems to be a coherent idea, it does not follow that in adult subjects' minds such an exchange can easily happen. For unstructured representations UR1 and UR2 to exchange causal roles all the causal connections UR1 has to other states and behaviour need to be assumed by UR2 and vice versa. That is, for minds in which a large number of learned connections are already firmly in place actual causal role exchange seems to be impossible. Therefore, what the possibility of causal role exchange really means is this: the role an unstructured representation actually plays in a cognitive system could have been filled by some other unstructured representation equally well, in other words, the unstructured representation in question does not fill the role it actually fills necessarily.

This is an important consequence of being unstructured since it entails the *functional un-analysability* of unstructured representations. Since an unstructured representation does not fill the role it actually fills necessarily, knowing what role it fills (no matter how detailed the description is) does not help in specifying what unstructured representation it is that actually fills the role. Although the causal role actually filled by the unstructured representation is characteristic of the unstructured representation as a part of the actual system, since the very same causal role could have been filled by a different unstructured representation, it does not characterise the actual filler *uniquely* and thus does not distinguish it from other possible fillers.[27]

5.2.3 Explaining the epistemic gap within the Fodorian framework

With all the resources at hand, it is time to see how monadic markers embedded in a Fodorian framework are able to account for the occurrence of the epistemic gap.

First, recall the case of the ornithologist who has never ever seen a day-old chick.

[27] Here my main aim was only to flash the idea that functional un-analysability might be accounted for in terms of certain characteristics of unstructured representations. For a detailed discussion of how features of unstructured representations entail functional un-analysability, see §5.3.2.

Though the A-slot of the file associated to her DAY-OLD CHICK concept is filled with detailed propositional knowledge about day-old chicks, its P-slot is empty. Now consider that this is the first time she tries to imagine a day-old chick. Based on descriptions mentioning such simple shapes (e.g. feather, wing, claw, etc.) and colours (e.g. yellow) which the ornithologist has prior experience of she is able to simulate the look of a day-old chick. That is, her abstract knowledge drives her imagination resulting in a perceptual template which now fills the so far empty P-slot. Note that the P-slots of the files associated with the concepts YELLOW, FEATHER, and so on, must have been filled in order for abstract knowledge to be able to generate the day-old chick template by simulating the look.

Now compare this case with the case of Mary (Jackson, 1982, 1986) the future neuroscientist, who has never ever seen anything red.[28] However, she knows everything there is to know—in terms of descriptions—about seeing something red. That is, just like in the case of the ornithologist, though the A-slot of the file associated to Mary's concept RED is filled with detailed propositional knowledge, its P-slot is empty. Nonetheless, there is an important difference between Mary and the ornithologist, namely that whereas the ornithologist is able to fill the relevant P-slot with a day-old chick template by simulating its look based on constituent structure, Mary (before leaving her room) is unable to fill the P-slot of the file associated to the concept RED with a red experience template. What makes the difference here is the fact that while the perceptual representation corresponding to the experience of seeing a day-old chick is structured the perceptual representation corresponding to the experience of seeing something red is not. The latter is a monadic marker, an unstructured representation, and as such its perceptual template cannot be simulated on the basis of the abstract knowledge Mary could acquire within her chamber.[29]

To see why this is so, let's first unpack what happens with Mary. Before her release,

[28] Cf. §2.2.4.

[29] Perceptual templates might best be thought of as patterns of activation within the perceptual system, i.e. as stored perceptual representations. Cf. §3.5 and Footnote 4 in §5.1.2.

she learns from books that seeing something red is identical with a salient neurological response pattern in subjects' V4 what she calls 'neuro response X', and so forms the concept NEURO RESPONSE X of it. That is, the A-slot of the file associated to pre-release Mary's NEURO RESPONSE X concept is filled with relevant neurological information, but its P-slot is empty—she has no idea of the relevant phenomenal character of occurrent neuro responses.[30]

Then, after her release, Mary is shown a piece of paper with a patch of red on it (Nida-Rümelin, 1996). She locks the dummy concept XYZ to this kind of stimulus.[31] Whereas the P-slot of the associated file is filled with the perceptual template of red, its A-slot is at least nearly empty. Nonetheless, deployment of the concept XYZ helps Mary in recognising and categorising new varieties of the red stimulus, and in this way, contextual information can fill the A-slot of the associated file with information like 'it is the colour of London buses', or 'it is the colour of ripe tomatoes'.[32]

Note, that there is no *a priori* connection between Mary's XYZ and NEURO RESPONSE X concepts. First, Mary cannot fill the A-slot of XYZ with information like 'neuronal activation X' based on introspection when seeing something red, because this level of processing is encapsulated—inaccessible to conscious reflection.

Second, Mary is unable to fill the P-slot of NEURO RESPONSE X with a perceptual template of a red experience solely on the basis of the content of its A-slot. Since the

[30] Given that Mary knows everything there is to know about the nervous system, and that the only kind of visual stimuli she lacks is colour-stimuli, it is likely that the P-slot of the file associated to her NEURO RESPONSE X concept will not be empty but filled with a perceptual template of an assembly of activated neurons. Nevertheless, it won't help her in simulating the look of red.

[31] According to Fodor's innateness thesis, our conceptual system is full of dummy concepts waiting to get locked to a new kind of stimulus. One does not need to buy innateness here though, for one could argue that Mary acquires a new concept XYZ which is nomically locked to perceived redness. For example, this new concept could be a newly acquired perceptual concept in Papineau's (2007) sense—cf. §3.5.

[32] Again, compare this with how perceptual concepts work in Papineau's (2007) account—cf. §3.5.

perceptual representation corresponding to the red experience is a monadic marker, it has no constituent structure that could help Mary (like prior experiences of wings and claws helped the ornithologist). Nor has Mary access to the features[33] of red experiences—features of monadic markers do not form standalone experiences, the only way to experience them is as features of the overall experience (seeing something red), which is exactly what Mary lacks.

Moreover, neither structural nor functional information conveyed by the abstract knowledge in the A-slot is of any use. Structural information cannot help in simulating monadic markers due to the fact that monadic markers lack any structure, and functional information cannot help either due to the functional un-analysability of monadic markers.

The only way for Mary to connect the two concepts is via filling the A-slot of the XYZ concept with contextual information she is familiar with from her pre-release studies. If, for example, Mary learns that XYZ is utilised when seeing the colour of London buses, and if she has learned inside her room that neuro response X answers to the colour of London buses in neuro-typical individuals, then she can conclude that the two concepts XYZ and NEURO RESPONSE X co-refer. In effect, the A-slots of the two concepts can get merged—all the conceptual information associated with XYZ gets associated with NEURO RESPONSE X, and vice versa: all scientific information associated with NEURO RESPONSE X gets associated with XYZ.

However, even if one can link and merge the A-slots, the phenomenology is tied to

[33] Hue, saturation and lightness. Cf. §4.1.1.

the P-slots.[34] This entails, that when the content of a P-slot is based on monadic markers, abstract knowledge necessarily leaves out what it is like to have the relevant experience—since such a P-slot cannot be filled solely on the basis of the corresponding A-slot. That is, no matter how detailed abstract knowledge one can get via merging the A-slots of scientific concepts, it remains ineffective in simulating the phenomenology if one lacks the basic constituents—monadic markers—of the perceptual template.

This is the epistemic gap, and also the reason why we have the intuition that the physical and the phenomenal are distinct. Note that the explanation introduced above does not need to presuppose any difference at the conceptual level. So-called phenomenal and physical concepts share the same characteristics inherited from the Fodorian framework: they are atomic LOT symbols with associated files containing abstract information and perceptual templates. The explanatory work is done by the fact that monadic markers form the representational bases of perceptual templates alone. This, then, validates the fundamental tenet of the Monadic Marker Account, claiming that the right level of analysis is the level of perceptual representations.[35]

[34] The P-slot, as we have seen it, contains a perceptual template, which might best be thought of as an organised set or pattern of stored perceptual representations some of which are monadic markers (cf. §5.3.1). According to the Monadic Markers Account, the phenomenal qualities of a conscious experience are determined by the structural information extractable from this pattern of representations plus the way monadic markers are interpreted by central processes (cf. §4.3.3). This is why the phenomenology is tied to the P-slot. See Papineau (2007), who also argues that the phenomenology is tied to perceptual templates (cf. 3.5).

[35] Note that all that has been said so far is neutral with regard to Papineau's or Aydede and Güzeldere's claim that we deploy the same concepts in introspection and when referring to object stimuli. That is, I do not argue against their account of how phenomenal concepts work. What I am arguing for, however, is the claim that the level of phenomenal concepts is not the right level of analysis—the special features of phenomenal concepts (if any) can be accounted for in terms of the features of monadic markers, and moreover, the Monadic Marker Account is able to explain the philosophically most important features of conscious experience directly, i.e. without relying on any particular theory of concepts.

5.3 The Explanatory Power of the Monadic Marker Account

The previous two sections of this chapter have already flashed a few illustrations about the explanatory power of the Monadic Marker Account. §5.1 has clarified the relationship between the Monadic Marker Account and an influential approach derived from the Dretskean framework by Murat Aydede and Güven Güzeldere (2005), and John Kulvicki (2004, 2005). It has been shown that the Monadic Marker Account is compatible with information-theoretic psycho-semantics. §5.2 has further expanded the boundaries of the applicability of the Monadic Marker Account by showing how monadic markers can be fitted into a Fodorian framework. Moreover, we have seen how the existence of monadic markers accounts for the occurrence of the epistemic gap within these different psycho-semantic frameworks. In what follows, I try to demonstrate the real explanatory power of the Monadic Marker Account by showing how—without being committed to any particular psycho-semantic theory—it is able to account for such fundamental features of conscious experience as the facts that it resists functionalisation, and that it necessitates an epistemic gap. However, before turning our attention towards these issues, first let's consider how the Monadic Marker Account is able to make the apparent distinction between primary and secondary qualities intelligible, and thus supports the observations made in §4.1.2.

5.3.1 Explaining the primary-secondary quality distinction

In §4.1.2 it has been argued that the relationship between colour experiences and colour stimuli seems to be different from the relationship between shape experiences and shape stimuli. There is something more to colours than what is presented by colour experiences, whereas it seems not to be the case with regard to shapes (Kulvicki, 2005). Or to put it in another way: shape experiences seem to be *revelatory* in a way in which colour experiences are not (Jakab, 2003, 2006). Or to be put in yet another way: shape experiences seem to *resemble* the corresponding features of object stimuli, whereas colour experiences do not (Locke, 1690/1987).

The *locus classicus* of this revelation/resemblance claim is, of course, Locke's famous passage, where he formulates it in this way:

> "[T]he ideas of primary qualities of bodies, are resemblances of them, and their patterns do really exist in the bodies themselves; but the ideas, produced in us by these secondary qualities, have no resemblance of them at all." (Locke, 1690/1987, II, viii, 15)

That is, upon reflection, there seems to be a natural way in which the properties we experience as possessed by objects can be classified into two distinct sets. On the one hand, there are perceived properties which seem to reveal or resemble the corresponding properties of the object stimuli—these are the so-called primary properties. Typical examples are shapes, spatial structures in general, and motion. My experiences when looking at a day-old-chick inform me about the specific physical shape of the chick, and do so in a quite effective way in the following sense: 'shape science'—i.e. the scientific endeavour aiming at revealing the source of the information carried by my shape experience—up until a certain resolution, will not report anything very different from the shape as experienced. On the other hand, there are perceived properties which do not seem to reveal much about the corresponding properties of the object stimuli—these are the so-called secondary properties. Typical examples are colours, smells, tastes, sine-wave tones. My experiences when looking at the day-old-chick present the chick as being yellow. However, colour science reports that the source of the information carried by my yellow experience is a certain range of values of surface reflectance. Surface reflectances, as colour science reports them, seem to be very different from colour experiences.[36]

Observing these differences is one thing, explaining them and rendering them intelligible is another. The Monadic Marker Account is able help in just that. Remember, the Monadic Marker Account introduces the following picture. Only a

[36] Similarly, molecular shapes and vibrations seem to be very different from smells and tastes; air-compression waves seem to be very different from sounds, etc.

range of the whole representational hierarchy is accessible for central processes (in a relevant way). Central processes are able to extract structural information regarding the objects representations within this range stand for. However, due to the lack of central access to lower levels, representations at the earliest stage of this range are monadic markers. Simple experiences—i.e. the very colours, smells, tastes, and sine-wave tones, which are the paradigmatic examples of secondary qualities—correspond to monadic markers in the sense that the phenomenal character of these experiences is identical with how monadic markers are interpreted by central processes. Similarly, complex experiences—i.e. the very shapes and spatial structures, which are the paradigmatic examples of primary properties—correspond to higher level representations within the accessible range, and their structure corresponds to the structure central processes extract from the range of representations they access in the sense that the phenomenal character of complex experience is jointly determined by the structure extracted, and by how monadic markers are interpreted by central processes.

That is, there is a range of hierarchically organised representations such that representations at each level stand for constituents of what representations at the next higher level stand for. Though each representation is unstructured for the processes at the next higher level, central processes can extract structural information from the hierarchically organised range itself. This structure, together with how monadic markers at the bottom (i.e. entry) level of the accessible range are interpreted by central processes determine the phenomenal characteristics of complex experiences. In this sense, thus, monadic markers are the bases of complex experiences—they provide the bedrock for encoding further, structural features.

Monadic markers, as we have seen, screen-off the features of the represented objects —all they do is indicating or marking the occurrence of a certain state-of-affairs. They do not map the features of the objects they stand for, only indicate their presence. Nevertheless, the monadic markers together are able to carry certain structural information. The guiding analogy here is that of a set of tags. One can indicate quite complex states of affairs with certain arbitrarily chosen simple tags. In such a tagging system, even if the tags themselves have complex features they are interpreted as monadic wholes. Their individual features are unimportant, what is important is that they are reliably connected to appropriate states of affairs. In this way, they can be used as the basis of mapping ‘further’ structure. Once the tags are in place, their organisation can map features of how the individual states of affairs are related to each other. That is, with the whole set of tags it becomes possible to capture ‘inter-states-of-affairs’ structure. Contrary to individual tags, which do not reveal any structure about the states of affairs they stand for, the set of tags itself does reveal some structure: the structure characteristic of the interrelations of the states of affairs the individual tags represent.

The Monadic Marker Account, thus, tells us that our cognitive system tags the world with monadic markers. Certain states of affairs in the world (e.g. certain wavelength of light reflecting from a surface) are tagged with a particular monadic marker. The way the system interprets these monadic markers determines (is identical with) the corresponding conscious experience (e.g. perceived yellow). Since these tags are monadic they screen-off the features of the states of affairs they stand for, i.e. make them unavailable for central processes. This is why the corresponding experience will be very different from the feature it is an experience of. Nonetheless, other states of affairs—i.e. interrelations of those states of affairs of the world which are originally tagged (e.g. how different surfaces reflecting certain wavelengths of light are related to each other)—are registered not directly by monadic markers themselves, but by the organisation of the hierarchy of monadic markers and higher level representations also accessible for central processes. This whole range of representations is able to make information regarding the particular physical structure available for central processes. And since, in accordance with the Monadic Marker Account, the phenomenal character of the corresponding experiences is partly determined by the structure central processes are able to extract from the range of accessible representations, these experiences will be similar to the object features they are experiences of to a much greater extent than the experiences corresponding to monadic markers.[37]

That is, the Monadic Marker Account explains the primary-secondary quality distinction by claiming that those properties which are typically considered as secondary qualities are represented by monadic markers, whereas those properties which are typically considered as primary qualities are encoded by the organisation of higher level representations of the accessible range within the representational hierarchy[38] (cf. Kulvicki, 2004, 2005).[39]

[37] Note an interesting prediction of the Monadic Marker Account. Different modalities differ in the amount of structure discernible for subjects in their modality-specific conscious experiences. Similarly, the ability of discerning structure in one's experiences differ with expertise. The Monadic Marker Account predicts that this difference in the amount of detail one is able to discern in one's conscious experiences is due to a difference in the range of levels of perceptual representations accessible for central processes. For example, for untrained subjects, in the gustatory or olfactory systems all representations are monadic—only one level of the representational hierarchy is accessible for working memory. This range of accessible levels of representations, however, can be extended with expertise: the more expert someone becomes in mastering a certain modality, the more levels of the hierarchy of representations will be accessible for central processes.

[38] As a consequence of this, the Monadic Marker Account suggests that simple shape (edge) representations themselves are not monadic but are derived from a more primary set of representations (which, in turn, are the 'real' monadic markers of the system). For example, in the visual modality, edges might be derived as borders between regions of different colours (cf. Smith, 2010, who argues that we see the shape of an object in virtue of seeing its colours). Similarly, in the somatosensory system edges might be derived as borders between regions of different pressure-areas. Furthermore, shape-percepts generated by sonar-like sensory systems might also be derived from primary auditory tags, i.e. monadic sound representations. Note that this line of thought might be helpful in shedding new light on the multi-modal nature of shape experiences.

[39] John Kulvicki—using an information-theoretical terminology—argues for a similar explanation of the primary-secondary distinction. As he summarises his view: "perceptual systems must satisfy a certain structural constraint in order to make parts of their abundant information extractable and thus available to cognition. Specifically, the properties in virtue of which the system carries information about the environment [...] must be isomorphic [...] to the properties about which they carry information. [...] Perceptual representations of colors are not isomorphic to the colors *to the extent* that perceptual representations of shapes are isomorphic to the shapes." (Kulvicki, 2005, p. 105, emphasis added) The kind of isomorphism Kulvicki has in mind here is "constituent structure isomorphism [which] requires that whenever one information carrying property is a constituent of another, then the properties about which they carry information are also related in this way, and conversely" (Kulvicki, 2005, p. 113). Kulvicki motivates this distinction by a detailed account of how early stages of the visual system work (cf. Kulvicki, 2005, pp. 115-122). However, note a problem with this: as we have seen, the early stages of the visual system seem to be a bad candidate for being the 'locale' of conscious experiences of primary and secondary properties (cf. §4.3.2). Moreover, note a further problem with Kulvicki's approach: Kulvicki concludes that perceived colours are less isomorphic to colours than perceived shapes to shapes (cf. the emphasis added to the quote above). Contrary to this, however, Locke's original problem is not that perceived colours resemble the relevant features of object stimuli *less* than perceived shapes, but that they do not resemble them at all. The Monadic Marker Account, contra Kulvicki's account, explains why this is so.

5.3.2 Explaining functional un-analysability

Now let's move on towards the philosophically most important features of conscious experience: its functional un-analysability and the fact that it gives rise to an epistemic gap. §5.2.2 has already provided some intuition pumps supporting the claim that the Monadic Marker Account is able to explain why the phenomenal qualities of conscious experience resist functionalisation. The aim of this section is to develop a thorough argument supporting this conclusion.

Consider a machine consisting of a camera system with pattern recognition capabilities and an output unit, being able to distinguish 'joey-sightings' from all other forms of possible stimuli. When an object is presented to the camera, the camera system processes the information registered and if it identifies the stimulus as a joey it sends a specific electric signal to the output unit. The output unit produces a 'this is a joey' behaviour if it receives the specific electric signal—upon receiving any other kind of signal it stays in (returns into) a 'not joey' state.

In this example, the specific electric signal is an unstructured representation. For the output unit it screens off the features of the joey; the output unit has no access to such details as how the joey actually holds its tiny arms, or what state of development it is in. The specific electric signal carries the 'it's a joey' information without any further details about the object represented. Were distinct electric signals sent to the output unit corresponding to (carrying information about) e.g. cones, cylinders, beans, and their spatial positions (or a complex signal the parts of which encoded parts of the joeys), structural facts would not be screened off, and the output unit would have access to more details about the joey.[40]

The specific electric signal might have a lot of features (strength, duration, etc.), i.e. it might be a complex vehicle, still for the output system these features do not

[40] However, this still would be a system with monadic markers: in this case those signals would be monadic markers, which indicate the presence of cones, cylinders, etc. without carrying any more information about them. A system entirely without unstructured representations would be one that is capable of registering and processing all the details in the stimuli. Cf. §4.3.2.

represent independent features of the joey. In so far as this is the case—as the features of the electric signal do not convey independently identifiable bits of information for the output unit—it is an unstructured representations.

Note that as a mere marker or indicator of the presence of the joey, the features of the signal are insignificant. The actual features of the specific electric signal do not play any role in indicating 'joey-occurrences'. The system could have been set up in a quite different way, where a totally different electric signal would have played the very same role the actual signal plays. Compare this with the case where a complex signal is sent from the camera system to the output unit carrying all different sorts of information about cones, cylinders, spatial positions etc. Since different features of such a signal are interpreted independently by the output system, i.e. they are 'meaningful' for the rest of the system, they cannot be freely altered. The more information features of the signal carry about the features of the represented object the more difficult it is to change the actual signal to a different one. Similarly, the more features of the signal are processed independently by the output system the more tightly it is embedded into the system. That is, there are heavy constraints in play determining which complex signal could and which could not play a specific role in a given system.

However, in the case of unstructured representations, there are no such constraints. Mere indicators can freely exchange the role they play without an extensive restructuring of the whole system. Imagine a similar machine as the one above, except that its camera system responds to 'day-old chick occurrences' with a signal totally different from the one playing a part in the joey identifying machine. As long as the output units of the two machines interpret the incoming signal as an unstructured whole, the two machines could be reset in a way that the original joey signal plays the appropriate role in the day-old chick identifying machine and vice versa. Any kind of signal with whatever features is apt for playing the role of an unstructured representation in so far as the rest of the system treats it as an unstructured monadic whole.

Of course, once a system is set up, changing the mediating signal would ruin the desired behaviour. Nonetheless, the system could have been set up differently, with an entirely different unstructured representation filling the role. That is, unstructured representations in general can freely exchange the role they play in a system: the role an unstructured representation actually plays could have been filled by any other (arbitrary) unstructured representation. Unstructured representations, thus, are *functionally un-analysable*. The causal-functional role an unstructured representation plays in a given system does not specify the features of the actual role filler—it does not determine which unstructured representation it is that actually plays the role. Even if one knew everything there is to know about a system's organisation and the role filled by an unstructured representation within the system, one would still find it possible that the role might have been filled by another unstructured representation.

That is, if one accepts the Monadic Marker Account, i.e. if one accepts that monadic markers (which, as we have seen, are a special set of unstructured representations—cf. 4.3.2) as-interpreted-by-central-processes are identical with the phenomenal qualities of simple conscious experiences, then one becomes able to explain why phenomenal qualities resist functionalisation.

5.3.3 Explaining the epistemic gap

Besides functional un-analysability, the other (and closely related) fundamental feature of conscious experience is that it gives rise to an epistemic gap. As it has been introduced in §2.2.1, the claim that there is an epistemic gap between the physical and phenomenal realms amounts to the denial of an epistemic entailment from physical truths to phenomenal truths. In fact, the notion 'epistemic gap' is an umbrella term covering such manifestations as the explanatory gap (cf. §2.2.2), Mary's inability to deduce colour-facts from physical facts (cf. §2.2.4), and the conceivability of zombies (cf. §2.2.3). In this section, I argue that the Monadic Marker Account in itself is apt for explaining why the epistemic gap occurs. In order to support this claim, I show that any cognitive system with the right sort of representational hierarchy and cognitive architecture[41] will necessarily find physical explanations leaving phenomenal qualities out, phenomenal facts being non-deducible from physical facts, and zombies being conceivable.

According to the Monadic Marker Account, our phenomenal knowledge of complex experiences is based on our phenomenal knowledge of simple experiences, which in turn is of monadic marker states as they are interpreted by the central processes of the cognitive system. There are five different aspects of this situation scientific investigations (our physical knowledge) can inform us about: (1) what processes give rise to the monadic marker states, (2) what features their vehicles have, (3) what role monadic markers play within the system, (4) what objects they stand for, and (5) how they are embedded in the system. The source of the epistemic gap is that there is no *a priori* connection between any of these different bits of physical knowledge and how the system itself interprets monadic markers, simply because of *restrictions in access.* Restriction in access affects both directions. On the one hand, access to how the system interprets monadic markers is privileged—it is restricted to the system itself—whereas, on the other hand, the system lacks access to any of those features scientific investigations can provide information about. To see why, let's consider each aspect individually.

(1) Though scientific investigations can tell us subtle details about the nature of the brain processes giving rise to monadic markers, i.e. about the processes within the input mechanisms and at lower levels of the representational hierarchy resulting in monadic markers, the central system has no access to these activities. This follows from the Core Hypothesis claiming that levels of the representational hierarchy below the level of monadic markers are inaccessible for relevant central processes,

[41] Implementing the Core Hypothesis discussed in §4.2.3.

and from standard considerations regarding information encapsulation denying a rich and direct information flow between the input mechanisms and the central system.

(2) Scientific investigations can also reveal fine-grained information about the nature of the neural signals acting as the vehicles of monadic markers. However, as we have seen, the physical properties of the vehicles are unimportant. Recall the example of the smoke indicating the outcome of the papal conclave (cf. §4.3.1). Information about the neural states implementing monadic markers is analogous with information about, say, the molecular structure of the smoke signal. These are not the properties the system processing the signal is sensitive to, i.e. these are not information carrying properties of the signal. Monadic markers are special, since they do not have an information carrying property-structure mapping features of the represented objects —they are interpreted as monadic wholes. Thus the central system is insensitive (and in this sense do not have access) to the neural features in question here.

(3) Another kind of information that can be the subject of physical knowledge is a specification of the causal-functional roles monadic markers play in the system they are embedded in. However, as we have seen in §5.3.2, this is of no help in matching physical knowledge with knowledge of how monadic markers are interpreted by a system. Central processes, by interpreting monadic markers, only *tag* the objects monadic markers stand for. The actual nature of these tags are independent of the specific role they play in the system. The outcome of the papal conclave, as it happens, is indicated by a smoke signal, but it could have also been indicated by a buzzer, for instance (cf. §5.1.1), or any arbitrary tag. Monadic markers are functionally un-analysable: knowing what causal-functional roles they fill does not tell us anything about their nature.

(4) Similarly, and relatedly, physical knowledge about the objects the presence of which is indicated by monadic markers will not help either. Since monadic markers have no information carrying property-structure mapping the features of the

represented objects, the properties of these objects are screened off for the system. Central processes cannot extract information from monadic markers about the physical properties of the represented objects.

(5) Finally, the last aspect is information about how monadic markers are embedded into the system. What physical knowledge can tell us about this is structural and organisational information depicted from outside the system. These structural and organisational characteristics, just like the causal-functional features, are, however, independent of the way monadic markers are interpreted by central processes. The right sort of organisation might be necessary for a monadic marker to play a particular causal-functional role, it is, nevertheless, unrelated to *how* monadic markers are interpreted. So, for example, even if the relative position of the smoke signal to St Peter's Square is essential for it to be a signal of the outcome of the papal conclave (had the fire occurred above a different rooftop, people on St. Peter's Square would have called for fire fighters), this relative position is unrelated to the nature of the actual tag.[42]

Now, note that all that a physical explanation can rely on are exactly these properties of the cognitive system: what processes give rise to monadic marker states, what features their vehicles have, what role they play within the system, what objects they stand for, and how they are embedded in the system. Since there are no connections between these physical properties and the way a monadic marker is interpreted by the system itself, any system fulfilling the above characterisation (entertaining a hierarchy of perceptual representations with central access to a range of this hierarchy) will necessarily give rise to an explanatory gap: the physical explanation will leave out monadic-markers-as-interpreted-by-central-processes. No matter how detailed physical descriptions are given, the question why a certain monadic marker is interpreted in a particular way will nevertheless remain an unanswered question.

[42] Note that the organisation of the system monadic markers are embedded in does play an essential role in carrying information about features that are characteristic of assemblies of those objects monadic markers stand for—cf. 'structured representations' as discussed in §5.3.1.

This is how the Monadic Markers Account explains the occurrence of the explanatory gap (cf. §2.2.2).

Similarly, with Mary (cf. 2.2.4). All Mary can learn in her black-and-white room are facts about the physical properties mentioned above. Since the way monadic markers are interpreted by a system itself (i.e. from the inside) cannot be deduced from knowledge about these physical properties, Mary is unable to come to a conclusion regarding how her (or any other person's) cognitive system will interpret the monadic marker standing for e.g. redness—i.e. how the system will tag the physical property redness,—no matter how detailed knowledge she has about the way monadic markers are embedded in a system or the roles they play (etc.). When she finally leaves her room and is presented a patch of red on a sheet of paper her visual system processes the information, and generates certain perceptual representations, a range of which then is accessed by (or becomes accessible for) the central system. Only then will she be able to grasp *from the inside* how her system tags colours, i.e. interprets monadic markers standing for colours. However, this piece of new knowledge will not reveal a new fact of the world—all Mary learns is an old fact (having a monadic marker with certain functional roles embedded in a cognitive system in a certain way, being processed in a certain manner, etc.) in a new way: from within the very system in question.

Finally, given that any physical information a system can collect about its working mechanisms falls short in characterising the way it is for the system itself to deal with monadic markers, it is possible for the system to conceive of scenarios where the same monadic markers with exactly the same vehicles, embedding structure and processing mechanisms are nevertheless interpreted in a different way (spectrum inversion cases), or not interpreted at all (zombie cases). That is, the conceivability of zombies (cf. §2.2.3) is a natural consequence of having monadic markers in one's cognitive system. However, from this conceivability no possibility follows. Zombies are conceivable, because due to restrictions in access first person information gathered by a system cannot be matched onto third person information. Nevertheless, the first and the third person information is about the very same fact. Metaphysically, there is only one thing.

In general, since by assumption, phenomenal character is identical with how certain representational states (monadic markers) properly embedded in a system are accessed/treated/interpreted *by the system itself*, and the result of this interpretation, i.e. the nature of the tag the system uses to indicate the corresponding states of affairs, is independent of what processes give rise to monadic markers, what features their vehicles have, what role they play within the system, what objects they stand for, and how they are embedded in the system, third-person physical knowledge about this system will lack any kind of *a priori* connection to the phenomenal domain.

To wrap up: in this section it has been shown that the Monadic Marker Account provides resources to explain fundamental features of conscious experience in purely physical terms. The fundamental features of consciousness in question are its functional un-analysability and the fact that it gives rise to an epistemic gap. The Monadic Marker Account equates the phenomenal qualities of conscious experience with the way monadic markers are interpreted by central processes of a cognitive system. On the one hand, due to their unstructured nature, monadic markers can freely exchange their roles within a system, and thus they themselves are functionally un-analysable. On the other hand, third person physical knowledge consists of knowledge about what processes give rise to monadic marker states, what features their vehicles have, what role they play within a system, what objects they stand for, and how they are embedded in a system. None of these features have *a priori* connections with the way monadic markers are accessed (interpreted) by a system itself. Therefore any system embedding monadic markers is subject to an epistemic gap: it will find monadic-marker-interpretations-as-from-the-inside unexplainable in terms of, and underivable from physical knowledge.

Note that the Monadic Marker Account does not tell us why it is the case that the internal interpretation of a monadic marker comes with the specific phenomenal character it does, or why it comes with any at all. But that is all right—the Monadic Marker Account wants to acknowledge the presence of the epistemic gap. It starts from the very observation that there is such a gap, and its main motivation is to provide a physical explanation of why this gap occurs. That is, the Monadic Marker Account does not bridge the epistemic gap. On the contrary, it acknowledges it, and accounts for it in terms of certain features of the cognitive-representational system. Consequently, the Monadic Marker Account disarms the corresponding anti-physicalist arguments and provides strong support for physicalism.

Part 3

Reductive Explanation

Chapter 6:
Reduction, Reductive Explanation, and Identities

6.1 Reduction and Reductive Explanation

In the previous part of the dissertation, explanations of why consciousness resist functionalisation and why it gives rise to the explanatory gap have been introduced. In this final part, I would like to concentrate on the question whether these explanations are reductive or not. On the face of it, since the explanations as spelled out in §5.3 rely only on the features of the cognitive-representational system, they seem to be reductive. However, the explanations explicitly utilise the identity claim formulated in §4.3.3, and thereby directly invoke phenomena from the target domain (i.e. from the domain of the phenomenon to be explained), which suggests that they are not reductive. In order to be able to evaluate this question properly we need to be clear about what reductive explanation amounts to.

However, before turning our attention towards reductive explanation itself, first I would like to pause here for a second and clarify a terminological issue with regard to reduction. As it happens, different authors use the same terms for different purposes. For example, Tim Crane uses 'reduction' as a general term, denoting a relationship, which has an ontological and an explanatory aspect (Crane, 2001a), whereas e.g. for David Chalmers, the term 'reduction' signifies only the ontological aspect (Chalmers, 1996).

This ontological aspect of reduction is an identity claim. As Crane puts it:

> "[A] reduction (in Huw Price's phrase) 'identifies the entities of one domain with a subclass of entities of another'. Or, to put it another way: we start off with the 'target' entity, X, and find a reason for identifying X with Y. Our reduction tells us something we didn't know about X: that it is Y. Claims of

> reduction in this sense are identity claims [...] Understood ontologically, then, a reduction of A to B involves the claim that A = B." (Crane, 2001a, p. 54)

Classical scientific examples of reduction—such as the reduction of temperature to mean molecular kinetic energy, or of water to H_2O—all show that the concept of reduction has this strong ontological connotation: they are typically understood as claiming that temperature is identical with mean molecular kinetic energy, and water is identical with H_2O.

However, this ontological claim in itself cannot exhaust the idea of reduction. As Crane draws attention to it:

> "For identity is a symmetrical relation, but a reduction of A to B is not a reduction of B to A. And there are plenty of identity claims which are not reductions: it would be (at best) pointless to say that the discovery that Hesperus is Phosphorus is a reduction of Hesperus to Phosphorus. What reduction needs, in addition, is the idea that the 'reduced phenomenon' is made more comprehensible or intelligible by being shown to be identical with the 'reducing phenomenon'. We understand thermodynamical phenomena better when we are shown that they are (so the story goes) identical with certain kinds of mechanical activity. And we understand mental properties better when we are shown that they are (so the story goes) identical with physical properties of the brain." (Crane, 2001a, p. 54)

The idea here is that reduction cannot be (only) identity, because there are identity claims which are clearly not cases of reduction (we do not reduce Hesperus to Phosphorus, or Bob Dylan to Robert Allen Zimmerman), and, moreover, whereas identifying A with B is symmetrical (A is identical with B = B is identical with A) reducing A to B is not (A is reducible to B ≠ B is reducible to A). This further connotation of reduction, which goes beyond a mere identity claim, is an explanatory aspect. By showing that temperature reduces to mean molecular kinetic energy, or that water reduces to H_2O we gain a better understanding of the characteristics of temperature and water.

It is this explanatory aspect which carries the epistemological virtue of reduction. Whereas the ontological aspect expresses that certain terms of our theories refer to the same object, and thus tells us something about the world itself, the explanatory aspect contributes by showing how our knowledge of the reduced phenomenon is related to our knowledge of the reducing phenomenon, and thereby makes the reduced phenomenon more intelligible to us. Reducing temperature to mean molecular kinetic energy shows us how temperature-phenomena as described by thermodynamics fit into the world of molecules as described by statistical mechanics.

6.1.1 Reduction and physicalism

Crane calls the ontological aspect *ontological reduction* and the explanatory aspect *explanatory reduction*. Typical cases of scientific reductions (e.g. the reduction of water to H_2O, or that of temperature to mean molecular kinetic energy, etc.) are such that they implement both aspects. However, the two aspects do not necessarily go together. To see the full taxonomy, consider the four possible combinations of ontological and explanatory reduction: (FR) *full reduction*, where one has both ontological and explanatory reduction; (POR) *pure ontological reduction*, where there is ontological but not explanatory reduction; (PER) *pure explanatory reduction*, where one has explanatory but not ontological reduction; and (ANR) *absolutely no reduction*, where neither ontological nor explanatory reduction is available.

In §1.3.2 and §1.3.3 we have already seen that the notion of reduction plays an important part in differentiating between various versions of physicalism. There, following the literature, a distinction has been made only between reductive and non-reductive physicalism. However, now that we have all the different combinations of ontological and explanatory reduction at hand, we can enrich this picture, and investigate how (FR), (POR), (PER), and (ANR) fit into the physicalist worldview.

Physicalism, as defined in Chapter 1, is a metaphysical doctrine, claiming that the prima facie non-physical is metaphysically determined by the physical.[1] Identity, the fundamental claim of ontological reduction, is a version of metaphysical determination, which makes (FR) and (POR) straightforwardly compatible with physicalism. Since identity is not the only form of metaphysical determination, (PER) and (ANR) are also reconcilable with physicalism. In fact, there are actual physicalist positions held by certain authors which fall into these categories.

Full reduction is the case where the prima facie non-physical is identified with something physical (hence it is committed to metaphysical determination), and moreover, an explanation is provided making the prima facie non-physical intelligible. This, as we have already seen, is the case in standard scientific reductions.

Pure ontological reduction also identifies the prima facie non-physical with something physical (and thus it is also committed to metaphysical determination), however, it denies any explanatory link between the two. This, for example, is the case of Phenomenal Concept Strategy.[2] Phenomenal Concept Strategy claims that phenomenal concepts and the corresponding physical concepts refer to the same entities, nevertheless any kind of explanatory link is impossible due to the mutual conceptual irreducibility of phenomenal and physical concepts (cf. §3.1).

Physicalism plus pure explanatory reduction denies identity, but nevertheless maintains some other form of metaphysical determination, and moreover, claims that the prima facie non-physical can be made intelligible on the bases of the physical. This, for example, is the case of role functionalism. According to role functionalism, functional role properties are not identical with realiser properties, nevertheless, via

[1] Or, in accordance with the conclusion of §1.6.4, physicalism about a certain domain claims that the distinguishing features of the domain in question are metaphysically determined by a base not containing these features or their governing laws.

[2] Crane's original example for the case of the ontological aspect without the explanatory aspect is Davidson's anomalous monism (Davidson, 1970; Crane, 2001a).

clarifying how realiser properties are able to fill the very causal roles definitive of role properties, role functionalism provides the relevant explanatory link. And since any realiser property filling the relevant causal roles necessarily brings about the related role property, metaphysical determination is ensured.[3]

Physicalism plus absolutely no reduction is the case of pure metaphysical determination—a case where the prima facie non-physical is metaphysically determined by (but not identical with) the physical, and no explanatory link helping us understand the prima facie non-physical better on the grounds of the physical is available. This, for example, is the case of weak emergence. Proponents of weak emergence argue that certain—typically biological (Bedau, 1997), but sometimes chemical (McIntyre, 1999, 2007), and even macro-physical (Laughlin & Pines, 2000; Batterman, 2002)[4]—phenomena, though metaphysically determined by the (micro-) physical, are nevertheless in principle unexplainable in physical terms due to certain theoretical limitations.[5]

6.1.2 Reductive explanation

The explanatory aspect of any reductive endeavour aims at accounting for a target phenomenon in terms of a base phenomenon, thereby advancing our understanding of the target phenomenon. Though Crane (2001a) calls it *explanatory reduction*, in the literature it is most often called *reductive explanation* (Chalmers, 1996; Block & Stalnaker, 1999; Hill & McLaughlin, 1999; Chalmers & Jackson, 2001; Kim, 2005).

[3] See, however, Block (forthcoming) for an argument claiming that role functionalism and physicalism —as metaphysical doctrines—are incompatible.

[4] Under a particular reading. See §1.6.3 for an ontological interpretation of Laughlin and Pines (2000) and Batterman (2002).

[5] In the literature the term reduction is often restricted to the ontological aspect (cf. e.g. Chalmers, 1996; Papineau, 2002; Kim, 2005), which renders (FR) and (POR) versions of reductive physicalism, and (PER) and (ANR) versions of non-reductive physicalism. This, however, yields a quite unfortunate and terminologically confusing result: role functionalism, as a version of physicalism plus pure explanatory reduction becomes a version of non-reductive physicalism—that is, reductive and non-reductive (though in different sense) at the same time.

A reductive explanation is only one of the many possible ways a phenomenon might be explained. Historical, teleological and causal explanations are other forms of making a particular phenomenon more intelligible. Very roughly, a historical explanation explains the genesis of a phenomenon by enumerating a series of events in chronological order leading to the explanandum. A teleological explanation explains a phenomenon via clarifying how its features qualify as appropriate means towards a specific end. A causal explanation contributes to our understanding of a certain phenomenon by revealing the chain of causes and effects resulting in the particular phenomenon.

Contrary to all of these versions of explanation, which are diachronic in nature, reductive explanation is synchronic. Instead of providing a chronological order of a temporal sequence of events, it relies on a set of phenomena as the explanans, which co-occur with the phenomenon to be explained. Moreover, as opposed to causal explanations, which are always intra-level (cf. Craver & Bechtel, 2007), reductive explanations are typically inter-level[6]—they advance our understanding of a higher level phenomenon by accounting for it in terms of lower-level phenomena. That is, reductive explanations are synchronic, and rely on lower levels as explanatory resources.[7]

A paradigm example of reductive explanation, for instance, is the explanation why water boils (or freezes) at a certain temperature.[8] Joseph Levine famously relies on this very example in his *On leaving out what it's like* (Levine, 1993). He provides the following sketch of how the specific reductive explanation in question works.

[6] Only typically, because certain models of reduction incorporate theory-succession where a phenomenon is explained not on the basis of a theory describing lower level phenomena, but on the basis of a newer theory describing the same level phenomena differently. Cf. Bickle (1998).

[7] How exactly levels are to be understood depends on the actual model of reductive explanation. In §6.2 I will distinguish between three different models of reductive explanation and investigate how they define higher and lower levels.

[8] Note that the facts that water boils at 100°C and that it freezes at 0°C at sea level are *a priori* truths, since the Celsius scale has been defined in accordance with the boiling and freezing point of water (cf. Kripke, 1980). Nonetheless, this problem does not arise if we set up the question by relying on the Kelvin or Fahrenheit scales.

> “Molecules of H_2O move about at various speeds. Some fast-moving molecules that happen to be near the surface of the liquid have sufficient kinetic energy to escape the intermolecular attractive forces that keep the liquid intact. These molecules enter the atmosphere. That’s evaporation. The precise value of the intermolecular attractive forces of H_2O molecules determines the vapour pressure of liquid masses of H_2O, the pressure exerted by molecules attempting to escape into saturated air. As the average kinetic energy of the molecules increases, so does the vapour pressure. When the vapour pressure reaches the point where it is equal to atmospheric pressure, large bubbles form within the liquid and burst forth at the liquid’s surface. The water boils.” (Levine, 1993, p. 129)

The passage explains why water boils by relying on the behaviour and governing laws of H_2O molecules. This reductive explanation is synchronic, since the molecular processes it cites are passing off at the same time when water boils, and it relies on lower levels since individual H_2O molecules are the constituents of a body of water. Once we understand what is going on at the lower level, the higher level phenomenon immediately becomes intelligible. In a very restricted sense, reductive explanation is eliminative: it eliminates any sense that there is anything extra going on in addition to the processes present at the lower level. Of course, reductive explanation is not eliminative in the sense of eliminative physicalism (cf. §1.3.1)—it does not eliminate the higher level phenomenon; a reductive explanation of water-properties in terms of H_2O does not render the term ‘water’ obsolete, on the contrary, it, in a sense, *justifies* our use of this term by showing how water-phenomena fits into the world of molecules (cf. §1.3.2).

The key characteristic of reductive explanation is that it removes a mystery related to the higher level phenomenon (the explanandum). As David Chalmers puts is: “once we have told the lower-level story in enough detail, any sense of fundamental mystery goes away: the phenomena that needed to be explained have been explained” (Chalmers, 1996, p. 42).

So reductive explanation is a synchronic bottom-up process connecting a lower level to a higher level and thereby making the higher level intelligible. The interesting question related to reductive explanation is concerned with *how it is able to connect the different levels*, that is, what resources it can rely on in order to be able to create an explanatory link between the entities and processes of two distinct levels. As Chalmers formulates it:

> "[I]n general, a reductive explanation of a phenomenon is accompanied by some rough-and-ready *analysis* of the phenomenon in question, whether implicit or explicit. [...] Without such an analysis, there would be no explanatory bridge from the lower-level physical facts to the phenomenon in question. With such an analysis in hand, all we need to do is to show how certain lower-level physical mechanisms allow the analysis to be satisfied, and an explanation will result." (Chalmers, 1996, pp. 43-44)

In the following section, I will introduce three major approaches to reductive explanation and investigate this very question—what kind of analysis or other resources allow them to make the explanatory leap from a lower level to a higher level.

6.2 Models of Reduction

As we have seen in §6.1 reductive explanation is in fact an important aspect of any general reductive attempt. In this section, I examine three different models of reduction, with special focus on their explanatory aspect, and consider how they perform with regard to connecting the level of the explanans with the level of the explanandum.

6.2.1 Nagelian theory reduction

In 1948 Carl Hempel and Paul Oppenheim published a paper addressing the problem of scientific explanation. They suggested that: "the question 'Why does the phenomenon occur?' is construed as meaning 'According to what general laws, and by virtue of what antecedent conditions does the phenomenon occur?'"(Hempel, 1965, p. 246) That is, to explain a phenomenon is to show that its description follows deductively from laws (laws of an actual theory) and antecedent conditions. Similarly, to explain a law is to show that it follows deductively from other laws (and appropriate auxiliary premises). This is the so-called *deductive-nomological model of explanation*.

It was Ernest Nagel, who by applying the deductive pattern of explanation to the history of science introduced the notion of *reduction*. As he put it: “Reduction [...] is the explanation of a theory or a set of experimental laws established in one area of inquiry, by a theory usually though not invariably formulated for some other domain” (E. Nagel, 1961, p. 338). According to Nagel, a fundamental feature of reduction is that a so-called primary science (the reducing theory or base theory) absorbs a relatively autonomous secondary science (the theory that is reduced, i.e. the target theory). That is, in the history of science there are more inclusive theories which replace antecedent (less inclusive) theories by absorbing their laws and covering their observational phenomena.

According to Nagel there are two types of this absorption or reduction. He distinguishes homogeneous reduction from heterogeneous reduction. In *homogeneous reduction* a phenomenon or law of a theory is incorporated into another theory which utilises “substantially the same” terms that occur in the first theory. In the second type of reduction—in the *heterogeneous* case—the incorporating theory lacks some of the terms in which the phenomena or laws of the theory to be incorporated are expressed. Nagel’s own example is the reduction of classical thermodynamics to statistical mechanics. Terms like ‘temperature’ and ‘entropy’ occur in the laws of classical thermodynamics but are not present among the terms of statistical mechanics (E. Nagel, 1961, pp. 339-345).

So for Nagelian reduction the different levels in question are levels of descriptions as represented by different scientific theories. In this sense, temperature-related phenomena are at the level of phenomenological thermodynamics, whereas molecular motion-related phenomena are at the level of statistical mechanics.[9] The problem of Nagel's heterogeneous case is that the vocabulary utilised to form the descriptions of the level of thermodynamics is different from the vocabulary utilised to form the descriptions of the level of statistical mechanics. The explanatory tool of the Nagelian approach—namely the deductive-nomological method—needs to bridge this gap between the level of the explanans and the level of the explanandum.

For this reason, in the heterogeneous case the process of reduction is not an obvious, self-evident process; one has to formulate some connections between the adequate terms of the different theories in question. Nagel emphasised that the conditions for reduction could be formulated only for branches of science that had been formalised. One requirement for formalisation is fixing the meanings of the terms occurring in the theories by rules of usage appropriate to each discipline. Given that this is the case, the following is a necessary condition for the reduction of a target-theory to a base-theory: for each term which occurs in the target-theory but not in the base-theory, there must be a connecting statement—a so-called *bridge law*—which links the term with an expression formulated within the base-theory.[10]

Before moving on, it is of crucial importance to understand correctly what Nagel meant by bridge laws. As we could see it explicitly in the second paragraph of this section, for Nagel reduction is explanation of one theory by another: one uses "descriptive terms" (E. Nagel, 1961, p. 339) formulated in the reducing base theory to express statements formulated in the target theory. What bridge laws are designed to do is connecting the descriptive terms of the base theory to those descriptive terms of the target theory, which are absent in the base theory. The condition of such a

[9] So as far as the term 'level' is associated with an idea of hierarchical organisation, levels within the Nagelian framework are levels of the hierarchy of sciences. Compare this picture with that of Oppenheim and Putnam in their *The Unity of Science as a Working Hypothesis* (Oppenheim & Putnam, 1958).

[10] Note that 'bridge law' is not Nagel's term. Nagel calls what later reflections dubbed 'the requirement for bridge laws' as the "condition of connectability" (E. Nagel, 1961, p. 354).

connection is *co-reference*: Nagelian bridge laws pair theoretical expressions of different theories in such a way that the paired ones refer to the same state of affairs. This is how introducing a bridge law helps the base theory express a statement of the target theory: by connecting the theoretical term ‘temperature’ of thermodynamics to the theoretical expression ‘mean kinetic energy of molecules’ of statistical mechanics one is able to express facts about temperature within statistical mechanics *if* the theoretical expression ‘mean kinetic energy of molecules’ formulated within statistical mechanics picks out the same referent as the theoretical term ‘temperature’ of thermodynamics (cf. E. Nagel, 1961, p. 341). That is, original Nagelian bridge laws connect theoretical terms (or constructs of theoretical terms), and express co-reference.

Bridge laws are in the centre of Nagel’s approach. They provide the basic connections between the terms of the target theory and the base theory, and make the target derivable from the base. That is, these bridge laws are the resources Nagel relies on in order to bridge the gap between the level of the explanans and the level of the explanandum. Nagelian reduction is a logical derivation of the laws of the target theory from the laws of the base theory and some initial conditions—bridge laws. That is, bridge laws serve as initial conditions in the deductive process of Nagel’s approach.

The explanatory aspect of Nagelian reduction is manifested in the deductive-nomological method. Nagelian reduction makes the target-phenomenon intelligible on the basis of the base theory by proving that its description can be deduced from a set of premises consisting in (1) certain laws of the base theory, (2) some auxiliary premises specifying boundary conditions, and (3) bridge laws connecting the terms utilised in the description of the target-phenomenon with certain terms of the base theory. That is, Nagelian reduction explains the target phenomenon by showing that given these premises the description of the target-phenomenon necessarily follows.

The ontological aspect of Nagelian reduction is provided by bridge laws. As we have seen, by connecting terms of the target and the base theory, bridge laws express co-reference—by formulating a particular bridge law it is claimed that the target and base terms connected by the bridge law in question pick out the same thing. Creating a bridge law, for example, between 'temperature' and 'mean molecular kinetic energy' signifies the commitment that temperature as conceived by those thinking in terms of phenomenological thermodynamics and the mean kinetic energy of molecules as conceived by those thinking in terms of statistical mechanics are not two distinct entities—the two different uses of the two different terms refer to the same entity.

6.2.2 Hooker's reduction

Nagel's interpreters often refer to bridge laws as biconditionals (cf. e.g. Schaffner, 1969).[11] Biconditional bridge laws require that for each and every theoretical entity-type (those kinds picked out by the theoretical expressions) of the target theory there must be a nomologically coextensive theoretical entity-type of the reducing base theory. This very consequence is the target of the most common and severe critique of Nagelian reduction, the argument from multiple realisability, which claims that biconditional bridge laws are in general unavailable (see e.g. Putnam, 1967; Davidson, 1970; Fodor, 1974). This is why recent approaches to reduction try to do their best in order to construct a model of reduction, which is able to jettison bridge laws. As we will see in this and the next section, alternative approaches to reduction all advertise themselves with the claim that they are able to successfully evade bridge laws. Before turning our attention to Kim's functional model, which is the major contender of Nagel's approach in contemporary literature, first I would like to introduce a novel account of theory reduction, which, instead of offering an entirely different approach concentrates on refining the original Nagelian model.

[11] Note, however, that the original Nagelian model is not committed to biconditionals—bridge laws as one way conditionals would do the job perfectly. Nagel was well aware of this: in one of his footnotes he explicitly stated that bridge laws were not necessarily biconditional in form (E. Nagel, 1961, p. 355, n. 5).

This account of reduction is Clifford Hooker's model (Hooker, 1981), which has gained significant support in recent years (Bickle, 1998; Marras, 2002). According to Hooker's view—let T(B) and T(R) be the reducing base theory and the reduced target theory respectively—in order to reduce T(R) to T(B) one has to construct T(R) *, an analogue of T(R) in a way that T(B) together with some initial conditions C entails T(R)*. This T(R)* is the 'image' of T(R) within T(B)—i.e. T(R)* is entirely formulated within the vocabulary of T(B), and the relation between T(R)* and T(B) is straightforward deduction. Ausonio Marras summarises this reinterpretation of Nagel in three steps using the example of the derivation of Boyle-Charles' law, as part of the reduction of thermodynamics (*T*) to statistical mechanics (*T**):

> "1. The formulation of a number of limiting assumptions and initial conditions (*LA/IC*) centering around the identification of a fixed volume of an ideal gas with a fixed number of molecules.
> 2. The derivation, from the principles of statistical mechanics (*T**) together with *LA/IC*, of a law *L**, namely $pV = 2E/3$, which is the mechanical counterpart (an 'image' or 'close analogue') of the Boyle-Charles' law $pV = kT$ (call this *L*). *L** is of course entirely in the vocabulary of *T**.
> 3. The postulation of a bridge law (*BL*), $2E/3 = kT$, consequent upon a 'comparison' of *L** with *L*, enabling the formal derivation of *L* from *L**." (Marras, 2002, p. 238)

As opposed to Nagel's model, here a bridge law is not a premise of the deduction, but the consequence of a two step process of deducing an 'image' description within the base theory, and its comparison to the target description. So deduction remains the core concept of this kind of reduction too, but what gets deduced within Hooker's account is different from what gets deduced in the Nagelian approach. Nagel tries to deduce the original T(R) itself, and thus he needs inter-theoretical bridge laws, whereas Hooker deduces T(R)*, an 'image' of T(R), already within the vocabulary of the reducing theory T(B). By doing so, it seems as though Hooker's account succeeded in avoiding the need for bridge laws.

Besides the reduction relation, the other central notion of Hooker's model is the *analogue relation AR*, which connects T(R)* and T(R) and warrants that the reduction relation *R* holds between T(B) and T(R). As Hooker formulates it: "(T(B) & C ⊃ T(R)*) & (T(R)* *AR* T(R)) warrants (T(B) *R* T(R))" (Hooker, 1981, p. 49).[12] However, Hooker himself does not provide a general specification of these analogue relations. John Bickle (1998), in his *Psychoneural Reduction*, argues that although analogue relations are quite similar to Nagelian bridge laws, there is an important difference, namely that the elements of the analogue relations are only

> "ordered pairs of terms drawn from the nonlogical vocabularies of the two theories. Their sole function is to indicate the term substitutions in T(R)* that will yield the laws of T(R). […] No worry arises about the 'logical status' of ordered pairs of terms, even when one of a pair has no empirical extension. By themselves, these ordered pairs imply neither synonymy, synthetic identity, nor coextension." (Bickle, 1998, pp. 28-29)

Hooker's account tries to capture the whole range of possible reduction relations, with all the different kinds of ontological consequences. That is, the analogue relation is indeed a continuum. At one endpoint, one can talk about 'smooth reductions', where there are only a few and simple initial conditions which do *not* contain wildly counterfactual restricting conditions, and the laws of T(R) find close syntactic analogues within T(R)*. This is the case of inter-theoretic identities, where no large-scale corrections to the reduced theory are needed, and where the theoretical entities of the reduced theory are identified with (sets of) theoretical entities of the reducing theory. At the other endpoint there are 'bumpy reductions'—the initial conditions do contain numerous and wildly counterfactual restricting conditions, and even with the help of these extreme conditions the analogue relation between the laws of T(R) and T(R)* remains weak (Bickle, 1998, p. 29). The ontological consequence of 'bumpy reduction' is elimination, i.e. the replacement of the theoretical entities of the reduced theory by that of the reducing theory.

[12] In the Hooker quote 'C' stands for the same set of initial conditions which are designated by 'LA/ IC' in the Marras quote.

According to Bickel, we can place historical theory reductions along the continuum of the analogue relation. Near to the smooth or identity end there is the reduction of physical optics to Maxwell's electromagnetic theory. Near to the other end (the bumpy, or elimination end) there is the relation between phlogiston chemistry and oxygen chemistry. And in between, for example, nearer to the smooth end there is the reduction of Kepler's laws to Newtonian mechanics, and nearer to the bumpy end there is the reduction of classical equilibrium thermodynamics to statistical mechanics (Bickle, 1998, pp. 30-31).

That is, Hooker's model follows Nagel in that the explanatory aspect of reduction is manifested in the utilisation of the deducitive-nomological method, whereas the ontological aspect is provided not by bridge laws, but by the so-called analogue relation. It is the tool of analogue relation that bridges the gap between the levels of explanandum and explanans in Hooker's account. The analogue relation is designed to be much more 'flexible' than bridge laws: it requires neither identity nor coextension to be built into the premises of the derivation.

6.2.3 Functional reduction

Bridge laws as premises of a deduction are generally considered as a serious flaw of the original Nagelian model. For Jaegwon Kim, the main problem with the inclusion of bridge laws in the premises is that it prevents Nagelian reduction to implement real reductive explanation. For Kim, the source of the problem is that bridge laws connect the base entity with the target entity, and since in accordance with the original Nagelian picture they become part of the base theory, this brings the target entity down into the base domain, so the proposed explanation does not rely solely on base level resources (Kim, 2005, pp. 98-105).[13]

[13] David Chalmers shares this worry with Kim. As he puts it: "Perhaps we might get some kind of explanation by combining the underlying physical facts with certain further *bridging* principles that link the physical facts with consciousness, but this explanation will not be a reductive one. The very need for explicit bridging principles shows us that consciousness is not being explained reductively, but is being explained on its own terms." (Chalmers, 1996, p. 107).

Remember, reductive explanation ought to provide an account of a target phenomenon in terms of certain base phenomena, and it should rely, as explanatory resources, on the base phenomena only. As Kim puts it: “The explanatory premises of a reductive explanation of a phenomenon involving property F must not refer to F [...] or must not include a law pertaining to F” (Kim, 2005, p. 105). This requirement can even be strengthened by adding that the explanatory premises must not refer to any other property at the level of F—that is, they must refer only to properties at levels lower than that of F (Kim, 2005, p. 106).

Now a bridge law inclusive model of reduction clearly violates this requirement, since the very links it utilises to bridge the gap between the level of the explanandum and the level of the explanans work by evoking target phenomena in the premises of the deduction. However, the afore-mentioned gap between the different levels needs to be bridged, otherwise the explanation would not be able to “make a deductive transition from the base level, where our explanatory resources are located, to the higher level, where our explanandum is located” (Kim, 2005, p. 107).

Kim (1998, 2005), and others (Levine, 1983, 1993; Chalmers, 1996; Chalmers & Jackson, 2001) argue that this can be achieved by relying on conceptual connections, i.e. “definitions providing conceptual/semantic relations between the phenomena at the two levels” (Kim, 2005, p. 108). The desired conceptual connections are provided by functional analysis—by re-defining the target phenomenon in terms of its causal roles described in a base level vocabulary (Kim, 2005, p. 111).[14] The resulting model of reduction is the so-called functional model. Kim provides the following schematic form of functional reduction:

> “Step 1 [Functionalization of the target theory]
> Property M to be reduced is given a *functional definition* of the following form:

[14] What matters is that the relevant connection be *a priori.* See §6.3 and §7.1 for a detailed discussion.

> Having M = (def.) having some property or other P (in the reduction base domain) such that P performs causal task C. […]
>
> Step 2 [Identification of the realizers of M]
> Find the properties (or mechanisms) in the reduction base that perform the causal task C.
>
> Step 3 [Developing an Explanatory Theory]
> Construct a theory that explains how the realizers of M perform task C.”
> (Kim, 2005, pp. 101-102, original emphasis)[15]

Here, by ‘realizer’ of a functionally defined property M, Kim means any property in the base domain that fits the causal specification definitive of M.

To dig into the details, take A and B as the domains of the reduced and the reducing theories respectively, i.e. let A be the domain of the properties to be reduced and B the domain of properties serving as the reduction base. Property M ∈ A, and property P ∈ B.

Step 1 says that first a functional definition of M is given such as it ‘redefines’ or ‘reconstrues’ M in the vocabulary of the reduction base B: having M = having some P such that P performs C. Here the causal task C = {c1,…, ci, e1,…, ei}, where {c1, …, ci} are certain causes and {e1,…, ei} are certain effects (presuming that the same property can play a role in different cause-and-effect chains). The idea is that M is redefined as a functional property according to its causal roles.

[15] Versions of this schema can be found in the writings of different proponents of functional reduction. Kim’s functionalisation, for example, is *explication* in Chalmers’ terminology. The first two steps of Kim’s model correspond to explication, i.e. the clarification of what it is that needs to be explained by means of analysis, and explanation, where we see how that analysis comes to be satisfied by the low-level facts (Chalmers, 1996, p. 151).

Similarly, Levine refers to these two steps in the following way: “Stage 1 involves the (relatively? quasi?) a priori process of working the concept of the property to be reduced ‘into shape’ for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are.” (Levine, 1993, p. 132)

Step 2 is about finding a realizer property in the B base domain which satisfies the causal specification C, i.e. plays the specific roles defined by C such as {c1,…, ci} cause P to be instantiated and P causes {e1,…, ei} to be instantiated.

Step 3 says that a theory T(B) at the level of the B-domain is needed which can describe the causal relations {c1,…, ci, e1,…, ei} and P's part in them, and thus can explain how the realizer property P (as a theoretical entity of T(B)) fulfils the causal roles specified by C.

In this functional model of reduction there is no talk of bridge laws, and thus it seems (or so the proponents of this approach think) that, by using the functional model, one can get rid of all the problems related to the bridge-law-inclusive Nagelian model. The explanatory question asking why a property occurs at a given time can easily be answered: because having the property in question (the target property) is, by definition, having a property with a certain causal role, and the investigated system, at the given time, has a particular property realising the target property (i.e. filling the appropriate causal roles). That is, the target property is a functional property defined by its causal roles, so if in a given system a particular property realises these causal roles, then the system will automatically (by definition) exhibit the target property. The realizer property is exactly the property that fits the causal specification, so having the realizer property entails having the target property.

That is, the explanatory aspect of functional reduction is provided by a deductive argument, consisting in two sets of premises, and the conclusion. The first set of premises determines a theory of the base domain, i.e. a theory describing the behaviour of the base level phenomena. The second set of premises defines the target phenomena in terms of base level causal roles. From the first set of premises a causal description of base level phenomena follows. If one combines this with the definitions of the target phenomena in terms of base level causal roles, then the conclusion—the target phenomena—necessarily follows. Thereby, the target phenomena is made intelligible on the basis of the base phenomena.

In the functional model of reduction the definitions of the target phenomena (that is, their funcitonalisations) bridge the gap between the levels of the explanandum and the explanans. Kim argues that since defining higher level phenomena in terms of lower level causal terms is just a definition, the functional model does not violate the requirement that reductive explanation must not refer to target level entities. Definitions, Kim stresses, are not real explanatory premises, they are "cheap in proofs, they are free" (Kim, 2005, p. 111). It follows that, contrary to the Nagelian model, functional reduction yields true reductive explanation.

The ontological aspect of functional reduction (at least of the version proposed by Kim himself) is provided by an independent line of reasoning, the so-called causal exclusion argument, which is a version of the causal argument presented in §1.2. It identifies the higher level target phenomenon with the lower level base phenomenon. So Kim-style functional reduction, just like Nagelian reduction creates an ontological link between the phenomena of the reduced domain and the phenomena of the reducing domain in the form of an identity claim. However, a crucial difference between functional and Nagelian reduction is that whereas in the case of Nagel's model the identity plays a part in the deduction itself as a premise, and therefore it is a *prerequisite* of successful reduction, in the case of Kim's account the identity is rather the *conclusion* of the whole reductive process.

6.3 Identities and Reductive Explanation

In §6.2 I have introduced the received view with regard to the major approaches to reduction currently on the table, and explored their explanatory and ontological aspects. I will return to them in §7.2, and will criticise this received view. However, before doing that, first I would like to slow down and focus on the main topic of this whole chapter—the relationship between identities and reductive explanations. In this section I will investigate this relationship in more depth.

As we could see, classical Nagelian reduction needs identity claims as prerequisites, as premises in the deductive arguments fulfilling the explanatory requirement of reductive attempts. In stark contrast with this, modern approaches to reduction like Hooker's account or the functional model argue that identity claims, in fact, 'fall out' of the process of reduction, i.e. are provided as conclusions of the deductive arguments. In recent literature related to the debates over consciousness-based anti-physicalist arguments, the role identities play in reductive explanations became of central importance. In what follows, I introduce the opposing views.

6.3.1 Identities are justified by reductive explanation

Let's start by reconstructing one of the central examples of the literature—the way one can get from truths about H_2O to truths about water. The truths about water in question include, for example, the fact that water boils at 212 °F (or 373 K) at standard pressure (sea level) (cf. Levine, 1983, 1993), and the fact that water covers 60% of Earth (cf. Jackson, 1994, 1998, 2003).

Levine and Jackson (together with e.g. Chalmers, 1996; Chalmers & Jackson, 2001; Chalmers, 2003) argue that it is possible to get *a priori* from truths about H_2O to truths about water. Jackson (2003) refers to it as the "*a priori* passage principle", whereas Chalmers and Jackson (2001) calls it "transparent epistemic connection", "transparent entailment", "transparent explanation", and most expressively, "transparent reductive explanation"[16]. The argument runs as follows.

[16] Jackson defines the *a priori* passage principle in the following way: "for each true statement concerning our world, there is a statement in physical terms that *a priori* entails that statement" (Jackson, 2003, p. 84); whereas what Chalmers and Jackson have in mind under the name of transparent reductive explanation is something like that "an ideal reasoner who knows all the relevant physical facts and who has the concept 'water' would be in a position to judge that if the base facts are a certain way then the extension of 'water' is H_2O, and thereby be in a position to know various facts about water" (Polger, 2008, p. 111).

First, note that a sufficiently full and detailed set of base level (physical) theories and facts accounts for the fundamental properties like position, mass, etc., and hence the distribution and behaviour of H_2O molecules in the entire Universe. From this it is straightforwardly deducible[17] that 'H_2O covers 60% of planet Earth'. We would like to get from this to the conclusion that 'water covers 60% of planet Earth'. All physicalists agree that physical (in this case, H_2O) facts *entail* all other (in this case, water) facts; the question is whether (and how) H_2O facts *imply* water facts.[18] Jackson, and other proponents of the availability of an *a priori* passage accept that the straightforward move from 'H_2O covers 60% of planet Earth' to 'water covers 60% of planet Earth', although valid (in the sense that in every world where the H_2O claim is true, the water claim is also true), is only necessary *a posteriori*, not *a priori* (cf. Jackson, 1998, pp. 81-82). Nonetheless, they argue, the initial H_2O truth can be supplemented by further truths, which together *a priori* necessitate the desired water claim.

The required supplementation consists of two further premises. The first one is an *a priori* truth saying that 'water is the stuff that plays the water role' (or according to alternative formulations: 'water is the watery stuff'), where the 'water role' or 'watery stuff' is a shorthand for "being potable, odourless, falling from the sky, being the stuff that makes up various bodies of liquid of our acquaintance" (Jackson, 2003, p. 86), etc.—that is, is a shorthand for a list of the reference fixers for 'water'.[19] The second additional premise is the contingent *a posteriori* truth that 'H_2O is the stuff that plays the water-role'—which, as proponents of this view argue, is also deducible from the set of base level theories and facts. Supplementing the argument with this

[17] Supposing that it is true.

[18] "On our usage, *P* entails *Q* when the material conditional $P \supset Q$ is true: that is, when it is not the case that *P* is true and *Q* is false. An a priori entailment is just an a priori material conditional. For ease of usage, we will speak of a priori entailment as *implication*. On this usage, *P* implies *Q* when the material conditional $P \supset Q$ is a priori; that is, when it is possible to know that *P* entails *Q* with justification independent of experience. On this usage, entailment is a nonmodal notion, while implication involves an epistemic modality." (Chalmers & Jackson, 2001, p. 316)

[19] Given that 'watery stuff' is understood in this reference fixing sense, the *a priori* status of 'water is the watery stuff' "rests on the general thesis that 'N = the F' is *a priori* when 'F' specifies the reference fixers for 'N'" (Jackson, 2003, p. 86).

further empirical fact is compatible with the original claim, which states that there is *some* set of empirical facts concerning the base level which leads *a priori* to the target fact. That is, the argument runs as follows.

Transparent reductive explanation:
Premise 1 (deducible from the base set):
H_2O covers 60% of planet Earth.
Premise 2 (from *a priori* conceptual analysis):
Water is the stuff that plays the water role.
Premise 3 (deducible from the base set):
H_2O is the stuff that plays the water role.
Conclusion 1 (from Premise 2 and Premise 3):
H_2O is water.
Conclusion 2 (from Premise 1 and Conclusion 1):
Water covers 60% of planet Earth.

Since the conditional ‘Premise 1 & Premise 2 & Premise 3 → Conclusion 2’ is *a priori* knowable—so the argument goes—there is, in fact, an *a priori* passage from base level facts to target level facts. Moreover, this is a fully qualified reductive explanation, since the water fact is explained purely in terms of H_2O facts. Still moreover, in the course of this reductive explanation, the identity claim that ‘H_2O is water’ is a conclusion rather than a premise. That is, proponents of transparent reductive explanation—i.e. those arguing for the availability of an *a priori* passage from the base level to the target level—think of identity claims as being justified by, or being the results of successful reductive explanations.

Note, that this is the very thought underlying Joseph Levine’s explanatory gap argument. In §2.2.2 we have seen that Levine pinpoints a crucial dissimilarity between standard scientific identities and the identity claims related to phenomenal consciousness. He argues that whereas consciousness involving identity claims demand further explanations, standard scientific identities do not. The reason for this is that in the standard scientific case our knowledge of the low level processes readily makes the high level phenomenon intelligible, i.e. there is an *a priori* passage from low level processes to a high level phenomenon similar to the one above in the water case. As Levine puts it:

> 'our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role [...] exhaust our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand.' (Levine, 1983, p. 357)

That is, we are persuaded of the identity between heat and mean molecular kinetic energy by being shown how the motion of molecules can play the very causal role, which is definitive of heat. In other words, once we know enough about the behaviour of molecules—and possess the concept of heat—we are immediately in a position in which we are able to pronounce on the identity claim in question.

This process must be familiar by now: concluding on an identity claim as a result of a deductive argument starting from a premise determining the theory describing the behaviour of the base level phenomena, and another premise defining the target phenomenon in terms of base level causal roles is, in fact, Kim's functional model of reduction. In accordance with this, Kim (2005) characterises all the earlier proponents of the *a priori* passage approach as deploying versions of his functional model of reduction.[20]

In §6.1.2 we have seen that any reductive explanation requires some kind of a resource rendering it possible to make the explanatory leap from the lower level of

[20] Note an important difference, though. Whereas Kim is explicit about that the requirement of functionalisation must involve a re-description of the target phenomenon in terms of *base level* causal roles, the Chalmers-Jackson approach deploys a conceptual analysis of the folk (target) concept 'water' in terms of other folk (i.e. *target level*) concepts like 'potable, odourless, colourless', etc. Here my purpose is only to draw attention to this distinction. I will come back to this problem, and discuss it in more detail in §7.1.

the explanans to the higher level of the explanandum. In the case of the functional model of reduction, this link is provided by the conceptual (functional) analysis of the target phenomenon. This is step one in Kim's schema: the target phenomenon must be re-defined in terms of certain causal roles. Analogously, in the detailed example of the reductive explanation of water facts in terms of H_2O facts presented above, the required link is provided by Premise 2, i.e. a conceptual analysis resulting in a list of the reference fixers of 'water'. Therefore, for those who think of reductive explanation in terms of the functional model of reduction, and hence for proponents of the view that identity claims are results of reductive explanations, it is conceptual (functional) analysis that makes the explanatory leap from the base level to the target level possible (cf. Levine, 1983, 1993; Chalmers, 1996; Kim, 1998, 2005). That is, *a priori* conceptual analysis plays the very same role bridge laws play as part of the explanatory aspect of Nagelian reduction: it connects the base level (the level of the explanans) with the target level (the level of the explanandum). In a certain sense, thus, the debate turns on the very fact whether the actual bridge closing the gap between the levels of the explanans and the explanandum is *a priori* available or not. Proponents of the *a priori* passage approach naturally claim that this bridge is *a priori* available via conceptual analysis, whereas opponents argue that it is not.

6.3.2 Identities do not need explanation

Block and Stalnaker (1999) famously raise a series of objections arguing that the kind of conceptual analysis required by the proponents of the *a priori* passage view is unavailable.[21] As they say it:

> "What is offered is not an argument for this [that a priori conceptual analysis is required to close the explanatory gap], but *examples* that show that *if* a conceptual analysis of a certain kind were always available, then we could use these conceptual analyses to account for the necessary a posteriori truths of reductive explanation. We have no quarrel with this conditional. What we doubt is that these conceptual analyses are very often available." (Block & Stalnaker, 1999, p. 14)

Consequently, Block and Stalnaker think that reductive explanations are rarely transparent. They argue that instead of *a priori* available conceptual links, the gap

[21] Block and Stalnaker attack both Premise 2 and Premise 3 of the transparent reductive explanation argument. Against Premise 2, Block and Stalnaker argue that *a priori* conceptual analyses of folk concepts are neither necessary, nor sufficient. Take, for example, the folk concept of life. In order to provide a transparent reductive explanation of features of life in terms of physical (base level) theories and facts, an *a priori* conceptual analysis of 'life' is required analogous to Premise 2 stating that 'life is the stuff that plays the life role'. Remember, though, on the face of it, this statement is a tautology, and thus clearly *a priori*, in fact, it is a shorthand for a list of the causal roles all living things play, i.e. the reference fixers of 'life' or 'living', like get born, die, reproduce, locomote, digest, respire, etc. However, drawing the line between living and non-living beings is notoriously difficult, and hence coming up with a list of necessary and sufficient conditions seems to be hopeless.

The other main problem with the transparent reductive explanation argument Block and Stalnaker identify is the requirement of a 'uniqueness clause'. Premise 2 and Premise 3 won't result in Conclusion 1 ('H_2O = water') unless the water role gets amended by a uniqueness and an indexical criterion stating that there is only one '*unique* stuff that plays the water role *around here*'. The indexical criterion is required because we do not want to exclude the "possibility that there are other [watery] stuffs elsewhere that are unrelated to our applications of the concept of water" (Block & Stalnaker, 1999, p. 17). The uniqueness criterion guarantees that H_2O = water: without it, it might be the case that besides H_2O, there are other stuffs (e.g. ghost water) around here that play the water role. However, according to Block and Stalnaker, the problem with such an amendment is that it "throws doubt on the claim that [Premise 3] of the argument is a microphysical fact" (Block & Stalnaker, 1999, p. 17), i.e. that 'H_2O is the unique stuff that plays the water role around here' can be deduced from the set of base level theories and facts.

Block and Stalnaker's objections haven't remained unanswered, though: Chalmers and Jackson (2001) and Jackson (2003) have offered reasons to believe that Block and Stalnaker's criticisms are mistaken. Both papers argue that the required conceptual analysis in question is different from the traditional understanding of conceptual analysis aimed at finding necessary and sufficient applicability conditions. (See more on this crucial issue in §7.1.) Moreover, Jackson (2003, especially pp. 168-169) argues that the uniqueness objection is readily answered by the fact that a 'stop clause' needs to be built in into the *a priori* passage principle anyway. Cf. Footnote 29 in §2.3.

between the levels of the explanans and the explanandum are bridged by classical (i.e. Nagel-like) connecting principles available only *a posteriori*.

> "[M]any of the required bridge principles connecting folk with scientific vocabulary (such as that water is H_2O, and that boiling is the particular microphysical process that it is) will not be analytic definitions: very often, what is needed are the notorious necessary a posteriori truths." (Block & Stalnaker, 1999, p. 5)

The "notorious necessary a posteriori truths" in question are proper identity claims. Since we are not able to get from H_2O facts to water facts by *a priori* conceptual analysis, we need to connect the two levels by other means. A straightforward way to create the required connections is to identify entities and processes of the target level with entities and processes of the base level: e.g. identify water with H_2O, and boiling with the particular microphysical process that it is.[22] The resulting model of reductive explanation, which I will call *non-transparent*, differs from the transparent version in that instead of the extra supplement captured by Premise 2 and Premise 3 of the original transparent reductive explanation argument (resulting in the identity claims in question as a conclusion), it directly relies on the identity claims as premises of the deductive argument.

Non-transparent reductive explanation:
Premise 1 (deducible from the base set):
H_2O covers 60% of planet Earth.
Premise 2 (*a posteriori* truth; empirical hypothesis[23]):
H_2O is water.
Conclusion (from Premise 1 and Premise 2):
Water covers 60% of planet Earth.

[22] More accurately, what happens here is this. We connect the folk concept 'water' with the scientific concept 'H_2O' by claiming that the extension of 'water' is the same as the extension of 'H_2O'. Cf. §6.2.1.

[23] See §6.3.3 for Block and Stalnaker's reasons for accepting such identity claims as premises.

Therefore, identities play a very different role in non-transparent reductive explanations compared to their role in the transparent version. This different role, however,—namely that they become premises rather than conclusions—means that non-transparent reductive explanations themselves give no independent reason for thinking that the identities in question are actually true. Proponents of the *a priori* passage view argue that this is a fatal flaw: identities should not be unexplained or *brute*—without any reason for believing in them why would anyone hold them true?

Block and Stalnaker resist by claiming that their opponents are mistaken in thinking that identities should be explained—in fact, Block and Stalnaker argue, identities do not need explanation. They support this claim by drawing attention to cases of identity claims involving proper names, like 'Cicero is Tully', 'Bob Dylan is Robert Allen Zimmerman' or 'Mark Twain is Samuel Clemens'. Block and Stalnaker tell the following story to illuminate their point:

> "Suppose one group of historians of the distant future studies Mark Twain and another studies Samuel Clemens. They happen to sit at the same table at a meeting of the American Historical Association. A briefcase falls open, a list of the events in the life of Mark Twain tumbles out and is picked up by a student of the life of Samuel Clemens. 'My Lord,' he says, 'the events in the life of Mark Twain are exactly the same as the events in the life of Samuel Clemens. What could explain this amazing coincidence?' The answer, someone observes, is that Mark Twain = Samuel Clemens. Note that it makes sense to ask for an explanation of the correlation between the two sets of events. But it does not make the same kind of sense to ask for an explanation of the identity. Identities don't have explanations (though of course there are explanations of how the two terms can denote the same thing). The role of identities is to disallow some questions and allow others." (Block & Stalnaker, 1999, p. 24)

What this example sheds light on is the importance of correlations in the context of formulating identity claims. The crucial step in this story happens when someone realises that the list of events studied by the historians interested in the life of Samuel Clemens is the very same list of events which is studied by their colleagues interested in the life of Mark Twain. The question 'What could explain this amazing coincidence?' asks for an explanation of this correlation—it asks how it can be that

Samuel Clemens and Mark Twain lived such a similar life. The identity claim is the answer to this question: it says that, in fact, there is only one life here—Samuel Clemens and Mark Twain lived the very same life because they were the very same person.

Block and Stalnaker argue that it does not make sense to ask why Mark Twain is Samuel Clemens. After all, the information we have gained by learning that Mark Twain is identical with Samuel Clemens is exactly that there is only one person. Asking why Mark Twain is Samuel Clemens seems to amount to asking why this person is identical with herself—which, according to Block and Stalnaker, is not the most sensible question to be asked. As David Papineau formulates it: "[t]he point is that genuine identities need no explaining. If 'two' entities are one, then the one doesn't 'accompany' or 'give rise to' the other—it *is* the other. And if this is so then there is nothing to explain." (Papineau, 2002, p. 144)

So instead of asking for an explanation of the identity itself, one should aim at explaining the concomitant correlation. Asking for an explanation of why the two sets of events are so tightly correlated is the legitimate question here, and the identity claim delivers the answer to that question. Mark Twain and Samuel Clemens visit the same places, meet the same people, and react in the same way *because* they are the same person. That is, the identity claim itself, instead of requiring further explanation, *provides explanation*.

The solution Block and Stalnaker offer focuses on correlations. It claims that what really happens is that we observe a correlation between two sets of events and want to understand it. Postulating identities is the means to this end. However, note that on the face of it, this situation is quite different from the original problem of trying to account for a target phenomenon in terms of some base level facts and theories. In fact, it is a misdescription of the *a priori* passage view to say that it asks for an explanation of an identity. Proponents of transparent reductive explanation do not start with observing an identity and then asking for an explanation. All they emphasise is that the identity itself does not remain unexplained. As we have seen in §6.3.1, in the process of accounting for higher level phenomena in lower level terms, the identities ‘fall out’ automatically, and moreover, readily explained—all the reason why we should hold them true are readily at hand.[24] In a certain sense, thus proponents of the *a priori* passage view do agree with Block and Stalnaker that identities do not need any *extra* explanation: once all the base level facts are given (and we possess the higher level concept) we can ‘read off’ the identity claim. And this is what is important from the perspective of the *a priori* passage view. Chalmers and Jackson try to support this position with the following passage:

> “A subject who knows all the qualitative truths in question—physical, mental, social—and who possesses the concepts ‘Mark Twain’ and ‘Samuel Clemens’ will be in a position to deduce that the identity is true, even if the subject is initially ignorant of it. The subject will be in a position to know that there was an individual who was known to his parents as ‘Samuel Clemens’, who wrote books such as Huckleberry Finn and the like under the name ‘Mark Twain’, whose deeds were causally responsible for the current discussion involving ‘Mark Twain’ and involving ‘Samuel Clemens’, and so on. From all this information, the subject will be able easily to deduce that Mark Twain was Samuel Clemens, and the deduction will be a priori in the sense that it will not rely on any empirical information outside the information specified in the base. So this identity is not epistemically primitive.”
> (Chalmers & Jackson, 2001, p. 355)

Note, however, that the deduction of the ‘Mark Twain = Samuel Clemens’ identity above is significantly different from how transparent reductive explanation works. Transparent reductive explanation starts with a full description of the base level, and then proceeds by connecting it with the target level via *a priori* conceptual analysis. In the above example of Mark Twain and Samuel Clemens, though, it is not the case that one is provided with a detailed description of the life of Samuel Clemens, and then, on the bases of a conceptual analysis of the name ‘Mark Twain’ one recognises

[24] Cf. how Conclusion 1 follows straightforwardly from Premise 2 and Premise 3 in §6.3.1.

that Mark Twain = Samuel Clemens.[25] Rather, the identity claim is formulated on the basis of the fact that "the events in the life of Mark Twain are exactly the same as the events in the life of Samuel Clemens" (Block & Stalnaker, 1999, p. 24); that is, on the basis of the observation that there is a special correlation between Mark Twain's and Samuel Clemens' lives: the very same events occur in the apparently distinct lives. How can it be that two persons live the same life? The answer emerges quite naturally: the two persons must be the same. True, once one knows everything there is to know about the events in Mark Twain's life and in Samuel Clemens life, one will immediately be able to deduce the identity in question. However, this has nothing to do with the transparent model of reductive explanation—it is a consequence of the special fact that the two lives consist of the very same events.

In this sense, the 'Mark Twain and Samuel Clemens' case as presented by Block and Stalnaker is a quite poor analogue to the typical mind-body issue. In the latter case, the two sets of events in question (the mental, i.e. target level, and the neural, i.e. base level) are stunningly different. Actually, the very fact that apparently there is

[25] Proper names refer directly, and not via associated descriptions (cf. Kripke, 1980). David Papineau argues in detail that this feature of proper names makes identities like 'Samuel Clemens = Mark Twain' perfect analogues to mind-brain identities. Papineau thinks that we refer to our phenomenal qualities with so-called phenomenal concepts, the crucial feature of which is that they—just like proper names—pick out their referents directly, and not via an associated description (Papineau, 2002, cf. Chapter 3). Since neither proper names, nor phenomenal concepts have associated descriptions, no conceptual analysis could connect them to other concepts. According to Papineau, this difference in associated descriptions results in a difference between standard cases of scientific identities and the mind-brain case. As he puts it:

"[W]e can thus explain 'why this quantity is *temperature*', understanding this as the question of why it is raised by inputs of heat and causes heat sensations in humans, and we can explain 'why this discharge is *lightning*', in the sense of explaining why it is produced by thunderstorms and illuminates the sky. [...] But the scientific examples are not really explanations of identities. We aren't explaining why this liquid is *water*—that is, why it is *the liquid* which in this world plays the role of being colourless, odourless, and so on. This would be to explain why this liquid is itself, which would be misplaced. Rather, we are explaining why this liquid is colourless, odourless, and tasteless. We are explaining why it satisfies the descriptions with which it is pre-theoretically associated. This is a perfectly good thing to explain, and I allowed above that physics can explain such things. However this is not a matter of explaining an *identity*—of explaining why some entity is itself—but rather of explaining why some entity possesses certain further attributes." (Papineau, 2002, pp. 148-150, original emphases)

And since in the case of consciousness (and proper names) there are no such further attributes—due to the fact that phenomenal concepts (proper names) don't have associated descriptions,—no further explanations are required.

nothing similar in mental and physical (neurological) events is an important part of the mystery of the mind-body problem (cf. e.g. Block, 2007). Thus a somewhat better example could be the case of Superman and Clark Kent. Superman and Clark Kent have very different abilities and characteristics, they never appear at the same place, etc. If we suppose that all there is to know about Superman is about him *qua* Superman (and vice versa with Clark Kent), then in this case, asking ‘How could they be the same person?’ seems to be a valid question.

Note the difference with the Mark Twain case. There, if one knows everything there is to know about Mark Twain and Samuel Clemens, then, as Block and Stalnaker set up the example, simply on the basis of the fact that a significant amount of Mark Twain related information will literally be the same as what can be known of Samuel Clemens, one will be able to conclude on the identity without the need of any kind of further analysis. Contrary to this, in the Superman case as introduced above, even if one has a full description of Superman *qua* Superman and of Clark Kent *qua* Clark Kent, one will not be able to deduce that ‘Superman = Clark Kent’, simply because none of the Clark Kent attributes will be recognisable amongst the Superman attributes[26], and no further conceptual analysis could reveal more[27]. Thus if one is told that actually Superman is Clark Kent, one seems to be justified in being sceptical, and asking for an explanation of how this could be true. In this case, then,

[26] Of course, there is one important sign which could arouse suspicion, namely the fact that Superman is always absent when Clark Kent is present, and vice versa.

[27] Cf. Kripke (1980) and Footnote 25 above in §6.3.2.

one asks for an explanation of an identity claim between two entities *because their accessible features are different*.[28]

Of course, this Superman case is still not a perfect analogue. Its flaw is that there are no real correlations in it—there are no 'Clark Kent events' which really co-occur with certain 'Superman events'. The mind-brain case is more like a combination of the two proper name involving examples: very different target and base level events (as different as in the Superman case) correlate really tightly with each other (as tightly as in the Mark Twain case). This is not unprecedented, though. As we will see in §7.2, standard examples of scientific explanations all share these characteristics. But before getting there, let's first summarise where we are at the moment.

Recall the original question at hand. It asked how we could get from lower level truths (e.g. truths about H_2O) to higher level truths (e.g. truths about water); that is, it asked for a model of reductive explanation. Levine, Chalmers, Jackson, Kim and others argue that the correct model of reductive explanation is the transparent version as depicted in §6.3.1. This model requires *a priori* conceptual analysis, which, according to Block, Stalnaker, Hill, McLaughlin, Papineau and others, is usually unavailable. However, they claim that despite the general unavailability of *a priori* conceptual analysis, reductive explanation of the target phenomenon is still possible. The model of reductive explanation they provide is the non-transparent version as depicted here in §6.3.2. Both models rely on identity claims in order to be able to

[28] There is, however, a way to convince the sceptical: by helping her hide in the dressing room of Clark Kent. If there our sceptical subject saw Clark Kent changing his outfit, and leaving the room as Superman, then this experience would most probably convince her. What would happen in this case, is that we would create an epistemic situation where the sceptical subject would have a chance to anchor her 'Clark Kent' and 'Superman' tags to the very same object. The problem with the mind-brain case is that due to the impossibility of crossing the subjective-objective dividing line, such an epistemic situation in which the subject could co-anchor her phenomenal and physical tags is unavailable.

In §7.1 and §7.2 I will argue that the mind-brain case is not the only example with such characteristics. I will show that in this respect classical scientific identities are very similar to the mind-brain case. Typically, in the case of cross-level scientific identity claims, when at least one of the featuring properties or processes is such that it is inaccessible for our human senses, and the two sides of the identity are defined both theoretically and experimentally by different frameworks (i.e. their observation-conditions are different), then it is impossible to create an epistemic situation where the necessary co-anchoring could be done. Nonetheless, as we will see, there is a working model of reductive explanation even in these cases as well.

conclude on target level truths—however, whereas the transparent version gets them 'for free' via *a priori* conceptual analysis, the non-transparent version introduces them as premises coming from empirical considerations. Herein lies the important role of correlations. Proponents of the non-transparent version argue that typically one becomes able to bridge the gap between the target and the base levels, i.e. reductive explanation becomes possible, if there are interesting correlations between the two levels. On the bases of these correlations one formulates the identities in question, which, on the one hand, are justified by the fact that they provide explanations of the correlations observed, and, on the other hand, connect the two levels, transfer explanatory power from the base to the target level, and thereby ensure reductive explanation.[29]

6.3.3 Identities are justified by inference to the best explanation

Identities, according to the non-transparent version of reductive explanation, rather than being conclusions of deductive arguments, are justified on other grounds—typically by invoking the so-called principle of inference to the best explanation.

The principle of inference to the best explanation (cf. Harman, 1966) starts from the available evidences, i.e. the observational facts, and infers to the truth of that hypothesis, which best explains these facts. In this sense, inference to the best explanation is reversed thinking: whereas in the case of deduction, for example, one moves from certain premises to the conclusion which follows logically from those premises, in the case of inference to the best explanation one moves backwards—one starts with the outcome, and looks for that premise which would have resulted in such a conclusion. The problem of this method is that there can be many possible premises resulting in the same conclusion, i.e. there can be many possible explanations of the same available evidence. This is why the principle looks for the *best* explanation of the given observation—it claims that that particular hypothesis of

[29] In fact, these two claims belong to two different approaches. Compare Hill (1991) and McLaughlin (2001) with Block and Stalnaker (1999). See §6.3.3 for more details.

the many possible hypotheses should be accepted as true, which is able to account for the observation better than any of its alternatives.

As it happens, different proponents of the non-transparent version of reductive explanation emphasise different aspects of the deployment of the inference to the best explanation principle. Christopher Hill (1991) and Brian McLaughlin (2001, 2010), for example, focus on the observed correlation between the target and the base domains, and claim that identities are justified on the grounds that they provide the best explanation of these correlations. As McLaughlin puts it in relation with the mind-body issue:

> "It is fairly widely believed that felt bodily sensations (aches, pains, itches, tickles, throbs, cramps, chills, and the like) have physical or at least functional correlates. That is to say, the following thesis is fairly widely believed:
> *Correlation Thesis.* For every type of sensation state, S, there is a type of physical or functional state, P/F, such that it is nomologically necessary that for any being, x, x is in S if and only if x is in P/F.
> The Correlation Thesis, if it is true, would not, of course, solve the mind-body problem for sensations. For the Correlation Thesis is compatible with virtually every theory of mind: not only with noneliminativist materialism and functionalism, but also with Cartesian Dualism, dual-aspect theory, neutral monism, and panpsychism.
> Some philosophers, however, myself included, maintain that the following thesis would offer *the best explanation* of the correlation thesis:
> *Type Identity Thesis.* For every type of sensation state, S, there is a type of physical or functional state, P/F, such that S = P/F." (McLaughlin, 2001, p. 319, original emphases)

Here, the available evidence, i.e. the observational fact, McLaughlin considers is the correlation between sensations and physical-functional states. The so-called Correlation Thesis expresses this observation. This observation, however, as McLaughlin acknowledges is compatible with many different explanation: e.g. functionalism, Cartesian dualism, dual-aspect theory, type identity, etc. Among these possible explanations, McLaughlin claims, the type identity thesis is the best explanation of the observed correlation.

Other proponents of the non-transparent version of reductive explanation emphasise other aspects of the deployment of the inference to the best explanation principle. For example, Block and Stalnaker, point out that identities can also be justified on the grounds that they allow for explanations, which otherwise would be unavailable. As they formulate it:

> "Why do we suppose that heat = molecular kinetic energy? [...] Suppose that heat = molecular kinetic energy, pressure = molecular momentum transfer, and boiling = a certain kind of molecular motion. (We are alluding to an empirical identity claim, not the a priori behavioral analysis considered earlier.) Then we have an account of how heating water produces boiling. If we were to accept mere correlations instead of identities, we would only have an account of how something correlated with heating causes something correlated with boiling. Further, we may wish to know how it is that increasing the molecular kinetic energy of a packet of water causes boiling. Identities allow a transfer of explanatory and causal force not allowed by mere correlations. Assuming that heat = mke, that pressure = molecular momentum transfer, etc. allows us to explain facts that we could not otherwise explain. Thus, we are justified by the principle of inference to the best explanation in inferring that these identities are true." (Block & Stalnaker, 1999, p. 23)

This passage emphasises that identities make it possible to transfer explanatory and causal power from the base level to the target level. This is why identities are able to act as a true bridge between the two levels and allow us to account for target level phenomena in terms of base level processes. Identities—Premise 2 of the non-transparent reductive explanation argument in §6.3.2—connect the target level to the base level by tying co-referring terms and expressions together. They map the target level onto the base level along these co-referring terms. In other words, identities are tools for projecting base level descriptions onto the target level. In this sense, they transfer "explanatory and causal force".[30]

The above example evokes an argument similar to the one illustrating non-transparent reductive explanation in §6.3.2. To see this, let's specify the identities in

[30] That is, on Block and Stalnaker's view, identities best explain target level causal phenomena (given base level causal processes) rather than correlations between target level and base level events.

more detail. Let’s say that ‘boiling = molecular process MP’ and that ‘heating = increase of mean molecular kinetic energy’. Then from facts and theories of the base level (in this case, statistical mechanics) it is possible to deduce the claim that ‘increasing the mean molecular kinetic energy of a certain state of molecules causes molecular process MP’. Adding mere correlation claims to this picture saying that ‘boiling correlates with molecular process MP’ and ‘heating correlates with increasing mean molecular kinetic energy’ leads us only to the conclusion that ‘something correlating with heating causes something correlating with boiling’. However, if we add the identity claims specified above, we can conclude that ‘heating causes boiling’.

> **Premise 1** (deducible from the base set):
> *Increasing the mean molecular kinetic energy of a certain state of molecules causes molecular process MP.*
> **Premise 2** (empirical hypothesis):
> *Heating is increase of mean molecular kinetic energy.*
> **Premise 3** (empirical hypothesis):
> *Boiling is molecular process MP.*
> **Conclusion** (from Premise 1, Premise 2, and Premise 3):
> *Heating causes boiling.*

Here the competing hypotheses are the correlation claim and the identity claim. Whereas the correlation claim does not allow us to draw a conclusion with regard to the target domain, the identity claim does contribute to accounting for target level phenomena. That is, the identity claim is the one amongst the competing hypotheses, which explains target level phenomena better. So in accordance with the principle of inference of to the best explanation, Block and Stalnaker claim, the identities in question are justified.[31]

[31] For a more recent defence of this position, see Block (forthcoming).

Of course, proponents of the transparent version of reductive explanation criticise the justification of identities on the basis of the principle of inference to the best explanation. Jaegwon Kim (2005), for example, offers a series of objections targeting both the general method of inference to the best explanation and the particular role identities play in explanations. In what follows, I am going to focus on what he has to say about the central topic of this chapter, the role of identities.[32]

Kim's fundamental problem is that, according to him, referring to identities as providing best explanations is deeply mistaken, simply because identities are not explanatory devices. Kim supports this claim by three slightly different arguments.

First, Kim argues that the Hill and McLaughlin approach is bound to fail, since identities do not explain correlations—rather they eliminate them. On the one hand, by claiming that there is correlation between two sets of events, one automatically commits oneself to the further claim that the two sets of events are two *distinct* sets of events. That is, in the present context of correlating target and base level events, Hill's and McLaughlin's correlation thesis really tells us that the target level phenomena are something over-and-above the base level phenomena. On the other hand, in the case of formulating identities, one inherently commits oneself to the claim that the two sets of events identified are, in fact, the same event. That is, the very message of an identity claim is that there is only one thing, and so the target phenomena are nothing over-and-above the base phenomena. In other words, Kim claims, an identity does not explain a correlation, rather it shows that, in fact, there is nothing to be explained, which means that the principle of inference to the best

[32] Though see Footnote 34 below (at the end of §6.3.3) for a brief summary of his other points.

explanation is unavailable to Hill and McLaughlin, because no explanation can be such that if the explanans is true then the explanandum is false.[33]

Second, Kim draws attention to the fact, that in typical scientific practice correlations are never explained by identities. True, science often progresses by first observing certain correlations and then trying to identify possible explanations of that particular correlation. Moreover, even inference to the best explanation might play a role in choosing between the alternative explanations. However, identities never appear among the alternatives—in scientific practice identities are not considered as candidates for providing explanations of correlations. Typically, observed correlations are either explained by underlying common mechanisms (i.e. a single process at some lower level underlying both phenomena correlating), or by a common cause. Kim illustrates this point by the examples of the correlation between thermal and electrical conductivity, and the correlation between tidal movements and phases of moon (cf. Kim, 2005, p. 134). In the former case, the observed correlation is explained by the claim that "both types of conduction involve the movement of free electrons through the lattice structure of metals" (Kim, 2005, p. 134)—i.e. by pinpointing an underlying common mechanism. In the latter case, the observed correlation is explained by the claim that both tidal movements and phases of moon are "causal effects of the relative positions of the earth and the moon in relation to the sun" (Kim, 2005, p. 134)—i.e. by finding a common cause.

Finally, Kim points out that it is not surprising that identities are not utilised in real scientific cases as explanations of correlations, since literally, identities do not explain anything at all. To drive this point home, Kim invites us to consider the

[33] Note that Kim (2005) attributes the same position to Block and Stalnaker, and consequently argues that the Hill-McLaughlin and the Block-Stalnaker implementations of the principle of inference to the best explanation are incompatible. However, this can't be the case. As we have seen in §6.3.2 Block and Stalnaker are explicit about the importance of the observed correlations. According to them, these correlations, and not the identities themselves, are the real targets of the explanatory questions. As they write it: "[n]ote that it makes sense to ask for an explanation of the correlation between the two sets of events. But it does not make the same kind of sense to ask for an explanation of the identity" (Block & Stalnaker, 1999, p. 24). That is, Block and Stalnaker agree with Hill and McLaughlin that identities are the answers to these explanatory questions targeting the observed correlations. See also McLaughlin's response to Kim for a similar conclusion (McLaughlin, 2010).

alleged explanation of why Cicero is wise in terms of the fact that Tully is wise. Such an explanation proceeds from the first premise saying that 'Tully is wise' through the second premise saying that 'Tully is Cicero' to the conclusion that 'Therefore, Cicero is wise'. Kim observes that:

> "If anyone should offer this as an explanation of why Cicero is wise, we surely would not take it seriously. [...] [The identity] seems to do no explanatory work. If it does anything to move the inference along, it is by allowing us to *rewrite* the premise 'Tully is wise' as 'Cicero is wise,' by putting equals for equals (that is, via the substitutivity of identities). The fact represented by the first premise 'Tully is wise' is the very same fact as the fact represented by the conclusion 'Cicero is wise'; in moving from premise to conclusion, the same fact is redescribed. There is no movement here from one fact to another, something that surely must happen in a genuine explanatory argument. Identities seem best taken as mere rewrite rules in inferential contexts; they generate no explanatory connections between the explanandum and the phenomena invoked in the explanans; they seem not to have explanatory efficacy of their own." (Kim, 2005, p. 132, original emphasis)

Apparently, Kim is committed to the view here that an explanation ought to connect one fact with another. Since identities do not connect different facts, but rather, at best, tell us that different descriptions are descriptions of the same fact, all they are capable of is acting as a rewrite rule. In other words, identities themselves "do not play a role in generating explanations; they only allow us to redescribe facts" (Kim, 2005, p. 134). Hence, Kim concludes, identities do not have any explanatory efficacy on their own.

However, even if identities are not the right tools for *generating* explanations, Kim acknowledges that they can play an important role in explanations: they can, as he says, "*defend* or *justify* explanatory claims" (Kim, 2005, p. 136). What Kim has in mind is this. Let's suppose that we have a base level theory from which we can deduce a certain claim about a particular base level phenomenon. Without the help of identities that's all we can have: an explanatory claim about a particular base level phenomenon. Now, the nice thing about identities is that they allow us to make explanatory claims which we wouldn't be able to make if the identities weren't available. Say, there is an identity connecting the particular base level phenomenon in question to some target level phenomenon. Then, since the identity rewrites our explanatory claim about the base level phenomenon into an explanatory claim about a target level phenomenon we get a new explanatory claim, which can be defended or justified on the basis of this very process.

Note that this is the same feature of identities which is emphasised by Block and Stalnaker. As we have seen earlier on in this section, Block and Stalnaker justify identities on the basis of the fact that they are able to transfer explanatory and causal power from the base level to the target level. By rewriting or redescribing base level phenomena in terms of the target level, identities *project* the base level explanation onto the target level, thus contributing to the explanation of the target level phenomenon.

However, Kim nevertheless thinks that it is not enough for endowing identities with their own explanatory efficacy. Take, for instance, the example of non-transparent reductive explanation of this section, the derivation of 'heating causes boiling' from facts and theories of the base level. Kim claims that the "explanatory activity is over and done" (Kim, 2005, p. 145) right after we get the first premise (namely that 'increasing the mean molecular kinetic energy of a certain state of molecules causes molecular process MP'). That is, Kim argues, the real explanatory activity is the deduction of this claim from the base level theory. Once it is done all that happens in the argument is that this claim gets restated with the aid of the two identities (Premise 2: 'heating is increase of mean molecular kinetic energy', and Premise 3: 'boiling is molecular process MP'). This restatement (the conclusion, i.e. the claim that 'heating causes boiling'), though is a new explanatory claim, it is only a redescription of a phenomenon that has already been explained by the first premise. As Kim himself puts it:

> "The identities [...] serve only as rewrite rules, and they are not implicated in the explanatory activity. All the explaining represented in the derivation occurs

> within [the base level], and when we derive [Premise 1 from the base level] theory, we are doing some genuine explaining. And that is the only explaining involved here. The identities kick in only after the explaining is finished. True, these identities do have a role in the derivation of [‘heating causes boiling,’] but this is not an explanatory derivation; rather, it is a derivation in which we put ‘equals for equals,’ and thereby redescribe in folk vocabulary a phenomenon that has already been explained.” (Kim, 2005, pp. 145-146)

If so, then Block and Stalnaker cannot be right when they justify the identities in question on the basis of the principle of inference to the best explanation, because these identities are not part of the explanations involved. The principle of inference to the best explanation cannot justify the identities in question if they themselves do no explanatory work. Since these identities are “not implicated in explanations of the sorts Block and Stalnaker have in mind” they “cannot be the beneficiaries of inference to the best explanation” (Kim, 2005, p. 146).

To summarise, Kim thinks that both the Hill-McLaughlin and the Block-Stalnaker version of the inference to the best explanation approach fail—the former because identities do not explain correlations, the latter because identities do not play an explanatory role in the deductive arguments considered.[34]

6.3.4 Is the non-transparent model really a version of reductive explanation?

Of course, proponents of the inference to the best explanation approach do not accept Kim’s objections. To answer the first challenge, McLaughlin (2010) pinpoints that even if Kim is right that identities and correlations exclude each other, it is irrelevant to the inference to the best explanation argument he and Hill offer. For what drives

[34] Kim (2005) formulates some additional worries in general about the principle of inference to the best explanation. The most significant of these additional worries emphasises that inference to the best explanation is a version of inductive reasoning, and as such it must respect the principle of total evidence—i.e. the data, or evidence, to be explained must be all the data relevant to the issue at hand. Kim claims, that if one also considers—over and above psychophysical correlations—data like “presumptive authoritativeness and privacy of first-person access to one’s own mental states, the persistence condition of persons, the multiple physical realizability of mental properties, the possibility of qualia inversion, the possibility of ‘zombies,’ and the like” (Kim, 2005, p. 129) then identities might cease to appear to be the best of the available explanations.

Kim to the conclusion that identities and correlations are incompatible is—as we have seen—the observation that by claiming that two sets of events correlate one inherently commits oneself to the further claim that the two sets of events are distinct. However, Hill and McLaughlin do not commit themselves to this further claim since they do not even commit themselves to the correlational claim in this serious sense. All Hill and McLaughlin presuppose is the thesis that "for any type of state of phenomenal consciousness C there is a type of physical state P such that it is true and counterfactual supporting that a being is in C if and only if the being is in P" (McLaughlin, 2010, p. 267),[35] which is perfectly compatible with C being identical with P. So even if Kim were right that identities and correlations are incompatible, the only conclusion that really would follow from this would be that Hill and McLaughlin gave a misleading name to the thesis above (cf. McLaughlin, 2010, p. 271).

However, McLaughlin argues, Kim is actually wrong: identities and correlations do not exclude each other. In fact, identities do explain correlations. To see how this works McLaughlin asks us to consider a derivation of a correlational thesis from an identity originally put forward by Jared Bates (2009). The argument runs as follows:

P1: $\mathrm{M} = \mathrm{P}$. (assumption)
P2: $\mathrm{M}a$. (assumption)
C1: $\mathrm{P}a$. (from P1 and P2 by substitutivity rule)
P3: $\mathrm{P}a$. (assumption)
C2: $\mathrm{M}a$. (from P1 and P3 by substitutivity rule)
C3: $\mathrm{M}a \leftrightarrow \mathrm{P}a$. (from P2-C1 and P3-C2; biconditional proof)
C4: $\forall x(\mathrm{M}x \leftrightarrow \mathrm{P}x)$. (from C3; universal generalisation)
(cf. Bates, 2009, p. 322)

[35] Compare this with its earlier version that has already been cited: "for every type of sensation state, S, there is a type of physical or functional state, P/F, such that it is nomologically necessary that for any being, x, x is in S if and only if x is in P/F" (McLaughlin, 2001, p. 319).

Here the final conclusion (C4) is a universal correlational claim, and P1, P2 and P3 are the three independent premises, among which the identity statement (P1) is the only one carrying explanatory weight. That is, this deduction shows that it is possible to derive a correlational claim from an identity statement, and moreover, in the course of this derivation the identity is the only explanatory premise.

Moreover, even if Kim is right in his observation that the identity featuring in the 'explanation' of 'Cicero is wise' in terms of 'Tully is wise' (cf. §6.3.3) does not play any explanatory role, it doesn't follow that no identity plays any explanatory role at all. Kim's main reason for denying that the identity statement 'Cicero is Tully' plays any significant explanatory role here is that, as he puts it, it is only a rewrite rule. 'Tully is wise' is the very same fact as 'Cicero is wise', we only redescribe this fact with the aid of the identity. Since, according to Kim, an explanation should move us from one fact to a different fact, the deduction of 'Cicero is wise' from 'Tully is wise' via 'Cicero is Tully' is not an explanation. However, as Bates (2009) shows it, Kim's generalisation is too hasty. There are arguments featuring identity statements as premises which do move from one fact to another. Consider, for example, the argument from the premise 'Tully = Cicero' to the conclusion that "Therefore Tully is here iff Cicero is here." (Bates, 2009, p. 323). This argument clearly moves from one fact to a different fact, since "Tully's identity with Cicero" and "the constant joint appearance of Tully and Cicero" (Bates, 2009, p. 323) are two different facts, thus the argument satisfies Kim's requirement concerning explanations. Moreover, it is precisely the premise that explains the conclusion. "To put it simply, once we understand that Tully just is Cicero, it is no wonder that wherever Tully goes, there goes Cicero. In fact, no correct explanation of this fact could fail to mention that Tully and Cicero are one and the same." (Bates, 2009, p. 323) That is, the identity is the best explanation of the correlation.

This is still not the end of the story. As McLaughlin points out, Kim is even wrong in requiring that an explanation should connect two different facts. This condition might be valid on causal explanations, but it is definitely not on reductive explanations. The main difference between a causal and a reductive explanation is exactly that whereas the former explains via moving from one fact to another, the latter explains via re-descriptions. Remember to the very starting point of this chapter, where we saw there that there were many different versions of explanation (cf. §6.1.2). Causal explanation is certainly one of them, however, here in this chapter we are concerned with another kind of explanation, namely the one which tries to provide an explanation of a target phenomenon in terms of some synchronously co-occurring base phenomenon. Both the transparent and the non-transparent versions of reductive explanation proceed via identifying certain target phenomena with the appropriate base phenomena—that is, via re-descriptions. The crucial difference between the transparent and the non-transparent version is that the later one relies on *a posteriori* identities, i.e. identities which are cognitively informative. Kim is right: the initial premise and the final conclusion of the non-transparent reductive explanations both talk about the same fact. However, they talk about the same fact in different ways. Identities create the connections between these different ways of talking about the same fact. That is, literally, identities inform us that these different descriptions are of the same fact and moreover, about how it can be that they are of the same fact. Our understanding, thus, is advanced by the identities in question.

So, contra Kim, explanation in general is not necessarily achieved by moving from one fact to another. In certain cases (e.g. causal explanation) it is achieved in such a way, but in other cases (e.g. reductive explanation) it is not. Explanation is all about making something—described in a certain way—intelligible by invoking either other things or other descriptions. In this sense, explanations have an ontological and an epistemic aspect. The epistemic criteria for an explanation is that the explanatory premises need to epistemically imply the conclusion (cf. McLaughlin, 2010, p. 298). And identities do contribute to this epistemic aspect of explanations. As McLaughlin puts it:

> “[E]xplanation has an epistemic dimension. Facts, in Kim’s sense, explain and are explained only under descriptions or conceptualizations. (That is true even in the case of causal explanations.) [Identities deployed by the non-transparent model] do their explanatory work within the epistemic dimension of explanation. Only by taking into account the epistemic dimension can we capture the idea that explanations provide understanding, and give reasons for belief.” (McLaughlin, 2010, p. 298)

That is, the identity statements non-transparent reductive explanations rely on are real *explanatory* premises, and thus they are open to be justified by the principle of inference to the best explanation. So the non-transparent version seems to be a viable alternative to the transparent model of reductive explanations.

This leaves us with our last question. Let’s accept that the non-transparent model advocated by Block, Stalnaker, Hill, and McLaughlin does provide genuine explanations, i.e. that the transparent and the non-transparent versions are on a pair as far as explanations are concerned—but does the non-transparent model qualify as *reductive* explanation?

In a certain sense, of course, proponents of transparent reductive explanation are right when they claim that the non-transparent version does not explain the target phenomenon *purely in terms of the base level*, since it explicitly relies on identity statements as premises, which themselves refer to certain features of the target domain. In another sense, though, the explanation a non-transparent argument provides is *reductive explanation good enough*. Even if we cannot ‘read off’ the identities in question from a full description of the base level, and need to rely on empirical considerations, the identities themselves still express co-reference; that is, they allow us to see how a story told in the vocabulary of the base level covers phenomena from the target level. Remember, the transparent version of reductive explanation also needs identities in order to be able to project base level descriptions to the target level[36]. The only difference between the transparent and the non-transparent case is the source of the identity claim—whether it comes from *a priori* or *a posteriori* considerations. But even in the *a priori* case, one needs to possess and master target level concepts.[37] That is, even the transparent version of reductive explanation relies on a considerable amount of knowledge about the target phenomenon. So, it seems that even the transparent model *fails* to qualify as an explanation *purely* in terms of the base level (in the strictest sense of 'purely'). On the other hand, though, if we allow a certain amount of information about the target phenomenon—necessary for connecting the base and target levels—to be included in the deductive argument, then, it seems, the non-transparent version qualifies as reductive explanation just as well as the transparent version. After all, contrary to what opponents of the non-transparent version sometimes exaggeratively claim, it does not explain the higher level phenomenon in terms of the higher level phenomenon itself. In other words, it does not provide trivial explanations. On the contrary: it tells a story formulated purely in the vocabulary of the base level, and then shows how it can be projected to the target level. In this, the transparent and the non-transparent versions are very similar. Premise 1 of both arguments tell the base level story. Then Premise 2 of the non-transparent version, and Premise 2 together with Premise 3 (via Conclusion 1) of the transparent version project this story to the target level. Premise 1 is formulated purely in terms of the base level in both cases. Then, again, the 'projection devices'—the tools for bridging the gap between the base and the target levels—include information about the target level in both cases. Hence, the two arguments go together: if the transparent version qualifies as an

[36] To see this, consider the role of Conclusion 1 in reaching Conclusion 2 in the transparent argument in §6.3.1.

[37] Note how Kim's approach differs from Chalmers and Jackson's approach in this respect. For Chalmers and Jackson, the reasoner needs to master the concept picking out the target level phenomenon so that she will become able to identify the reference fixers of that concept (cf. Premise 2 in §6.3.1). Note that these reference fixers, as identified by Chalmers and Jackson (e.g. odourless, colourless, etc. for water), belong to the *target* level. Contrary to this, Kim's approach re-defines the target level phenomenon in terms of *base* level causal roles. On the face of it, by this move Kim manages to get rid of the burden of relying on the target level. However, as we will see in §6.4, this move is problematic.

instantiation of reductive explanation, so does the non-transparent version as well. The only difference is that they rely on different projection devices.

Remember, as we have seen in the introduction to this chapter, there are many different kinds of explanation. Here we are interested in those, which make the target phenomenon more intelligible on the basis of some underlying base level. This is what we call reductive explanation—it ought to provide an account of a target phenomenon in terms of certain base phenomena, and it should rely, as explanatory resources, only on the base phenomena.[38] The rationale behind this requirement is the following. Consider three different strategies all aiming at providing an explanation of the observation that certain target and base level phenomena co-occur. Let the first strategy be the method of transparent reductive explanation. This framework emphasises analytic definitions or *a priori* conceptual connections between the target and the base. Consequently, it becomes able to account for the target level phenomena in terms of a fixed base level ontology, and thus accomplishes the task of reductive explanation. The second strategy to be considered is ontological emergentism. Within this framework, trans-ordinal laws connect the base and the target levels telling us that whenever a certain base phenomenon occurs a particular target phenomenon also occurs (cf. §1.3.3). Such trans-ordinal laws, when included in an explanation of why the target phenomenon occurred extend the ontology of the base level on the basis of which one accounts for the target level phenomenon (literally, such arguments add the target phenomenon plus the corresponding trans-ordinal law to the base ontology). By this, ontological emergentism clearly violates the requirement of reductive explanation.[39] In fact, the

[38] Note that this requirement, as Kim originally formulated it was straightforwardly false. Kim said: "The explanatory premises of a reductive explanation of a phenomenon involving property F must not refer to F" (Kim, 2005, p. 105). However, since terms of the base level descriptions (initial premises) co-refer with terms of the target level descriptions (final conclusions), the explanatory premises of all reductive explanations necessarily refer to their particular target phenomena. To avoid this trivial failure, it is better to reformulate this constraint in the following way: "the explanatory premise of a reductive explanation of a phenomenon involving property F under the name or description α must not contain a use of α" (McLaughlin, 2010, p. 290).

[39] Cf. the importance of the base ontology in concluding on a domain-relative terminology in Chapter 1.

motivation behind this requirement is to exclude strategies, which explain a target phenomenon by first extending the base ontology with the very target in question, and thus, literally, solve the task by cheating. Now consider the third strategy: non-transparent reductive explanation. Non-transparent reductive explanation relies on identities as explanatory premises. These identities express co-reference, so they do not extend the base ontology in any sense. On the contrary, the non-transparent model is just as conservative relative to the base domain as the transparent model. The description it projects onto the target level is a description derived from the initial base domain. So there is no cheating on the ontological side. And as we have already seen it, there is no cheating on the epistemic side either—*a posteriori* identities are cognitively informative, and therefore the projection of the base level descriptions onto the target level via these *a posteriori* identities really do explain, and do so in terms of the base level. That is, the non-transparent model fully qualifies as reductive explanation.

Chapter 7:
Reductive Explanation and Prior Identities

7.1 Does Transparent Reductive Explanation Deliver?

Before moving on, note that proponents of the *a priori* passage view might not feel persuaded by what has been said in the closing sections of the last chapter (§6.3.4). They might want to object that in the preceding section I have overemphasised the similarities, and turned a blind eye to the main difference between the transparent and the non-transparent model. This main difference is, of course, the fact that whereas non-transparent reductive explanation relies on unexplained identities (in the sense that they cannot be 'read off' from a full description of the base), the transparent model is committed to *epistemically non-primitive* identities.[1] Chalmers and Jackson formulates it in the following way:

> "It is sometimes held that 'identities do not need to be explained' (for example, Papineau 1993). Block and Stalnaker say something similar ('Identities don't have explanations'). But this seems to conflate ontological and epistemological matters. Identities are ontologically primitive, but they are not epistemically primitive. Identities are typically implied by underlying truths that do not involve identities. [...] Once a subject knows all the truths about DNA and its role in reproduction and development, for example, the subject will be in a position to deduce that genes are DNA. So this identity is not epistemically primitive. Of course, just as with other truths involving macroscopic

[1] Note that on the basis of this, Chalmers and Jackson think that the identities deployed by non-transparent explanations are epistemically quite similar to trans-ordinal laws, which, contrary to what I have said in the previous section, disputes the reductive status of non-transparent explanations. As Chalmers and Jackson put it:

"[In the non-transparent case,] at best, there is an explanation in terms of physical processes plus psychophysical identities. And epistemically, the psychophysical identities play exactly the same role as psychophysical laws. They are inferred from regularities between brain processes and consciousness, in order to systematize and explain those regularities. And most importantly, the identities are not themselves explained, but are epistemically primitive. [...] Ontologically, these identities may differ from laws. But epistemically, they are just like laws. They are epistemically primitive psychophysical 'bridging' principles that are not themselves explained, but that combine with physical truths to explain phenomenal truths. An explanation of the phenomenal will have two epistemically irreducible components: a physical component and a psychophysical component. By calling the bridging principles identities rather than laws, this view may preserve the ontological structure of materialism. But the explanatory structure of this materialist view is just like the explanatory structure of property dualism." (Chalmers & Jackson, 2001, pp. 353-354)

> phenomena, subjects do not typically come to know these identities by deducing them from microscopic truths. But the identities are so deducible all the same, and their deducibility is what makes the phenomena in question reductively explainable."
> (Chalmers & Jackson, 2001, p. 354)

That is, according to Chalmers and Jackson, and as we have seen, according to the *a priori* passage view in general, in all the standard cases, there are reasons for formulating identity statements. A cogent reasoner is able to conclude on the identity of water and H_2O because she knows *a priori* that 'water is colourless, odourless, etc.' (i.e. 'water is the watery stuff'), and is able to deduce from a full description of the base level (again, *a priori*) that 'H_2O is colourless, odourless, etc.' (and similarly with, e.g. genes and DNA). That is, one concludes on an identity statement because one realises that the two entities in question, in fact, share their features.[2]

Proponents of the *a priori* passage might even dare to claim that their analysis applies to the very examples their opponents rely on. Consider, for instance, an example mentioned by Brian McLaughlin. McLaughlin, after arguing that identities and correlations are compatible, goes on and cites the following case to illustrate that, contra Kim, in actual scientific practice correlations are not explained only by common causes and common underlying processes, but in certain cases, by identities as well.

> "[T]here is a third way correlations are sometimes explained in science: by appeal to identity claims. When Maxwell's calculations showed that electromagnetic waves have the same speed in a vacuum as the known speed of light, he famously made 'the bold conjecture' that light waves = electromagnetic waves (Harman 1998; Maxwell 1973). Electromagnetic waves are refracted when going from one kind of material to another in a manner that depends on the refractive indices of the material. When it was established experimentally that light has the same refractive indices as electromagnetic radiation, this was taken to confirm Maxwell's bold conjecture. The hypothesis that light waves are electro-magnetic waves was invoked to explain why (1) electromagnetic waves and light waves occur in the same spatial regions at the

[2] Cf. the cases of 'Mark Twain and Samuel Clemens', and 'Superman and Clark Kent' in §6.3.2.

> same time, why (2) electromagnetic waves have the same speed in a vacuum as light waves, and why (3) the refractive indices in materials are exactly the same for light waves and electro-magnetic waves. This explanation, touted as one of the greatest achievements of classical physics, is an explanation by appeal to identity. Moreover, the identity claim was, arguably, taken to be justified by inference to the best explanation of the correlations in question." (McLaughlin, 2010, p. 282)

Here McLaughlin tries to show that the reductive explanation of some optical phenomena in terms of some electromagnetic phenomena follows the non-transparent pattern. The identity claim 'light waves = electromagnetic waves' is an empirical hypothesis, a *bold conjuncture*, which is justified on the bases of the principle of inference to the best explanation, given its ability to explain observations that light and electromagnetic waves co-occur, they have the same speed in vacuum, and also share refractive indices in different materials.

Note however, that proponents of the *a priori* passage view could point out that anyone, who is presented with a full description of electromagnetism, becomes able to deduce that 'electromagnetic waves have speed *c* in vacuum; the refractive indices for them in mediums $m_1, m_2, \dots , m_i$ are $n_1, n_2, \dots, n_i$ respectively, etc.'. Then given that one masters the concept of light, i.e. knows that 'light has speed *c* in vacuum; the refractive indices for it in mediums $m_1, m_2, \dots , m_i$ are $n_1, n_2, \dots, n_i$ respectively, etc.', one becomes able to conclude *a priori* that 'light is an electromagnetic wave'. In other words, it is the fact that light and electromagnetic waves share their features that lies at the very heart of this example, and motivates the formulation of the identity statement. Therefore, this identity is epistemically motivated, i.e. non-primitive.

Now the familiar argument can kick in: since the mind-body case cannot be motivated in a similar fashion, it is fundamentally different from typical cases of reductive explanation.

In the first part of this chapter my main aim is to show that this line of thought is mistaken. In order to see why, I will provide a detailed analysis of one of the prototypical examples of transparent reductive explanation—the reductive explanation of the boiling point of water in terms of a full description of a base (in this case, molecular) level.

7.1.1 From H_2O to water

As it happens, the way one can move from H_2O facts to water facts is the prime example in the debate over transparent reductive explanation (cf. Levine, 1983, 1993; Block & Stalnaker, 1999; Levine, 2001; Jackson, 2003; Polger, 2008). For instance, Joseph Levine famously relies on the reductive explanation of ‘water boils at 212 °F at sea level’ in terms of H_2O facts to illustrate that the explanatory gap is unique to phenomenal consciousness (Levine, 1993). That is, Levine argues, there is an *a priori* passage from H_2O facts to the fact that ‘water boils at 212 °F at sea level’. My aim in this section is to reconstruct this particular case of purported transparent reductive explanation in detail. At first pass, the deduction runs as follows (I follow the general structure of transparent reductive explanations as discussed in §6.3.1):

Transparent reductive explanation of ‘water boils’ – *first pass*
Premise 1 (deducible from the base set):
H_2O boils at 212 °F at sea level.
Premise 2 (from *a priori* conceptual analysis):
Water is the stuff that plays the water role.
Premise 3 (deducible from the base set):
H_2O is the stuff that plays the water role.
Conclusion 1 (from Premise 2 and Premise 3):
H_2O is water.
Conclusion 2 (from Premise 1 and Conclusion 1):
Water boils at 212 °F at sea level.

This deduction, however, is problematic. Note that Premise 1 should be deducible from the base set, i.e. from facts and theories of the molecular level (or the 'chemical theory of H_2O', as Levine (1993) calls it). However, as it happens, Premise 1 is not so deducible. The source of the problem is this: H_2O molecules do not boil. The concept of 'boiling' is not part of the vocabulary utilised at the molecular level—it is a folk concept[3], belonging to the same vocabulary as 'water'. The claim that *can* be deduced from the base set is something like the following: 'H_2O is engaged in behaviour B[4] at 212 °F at sea level'. So for the deductive argument to succeed it needs to establish a connection between 'boiling' and 'molecular behaviour B'. As we have seen, transparent reductive explanations typically establish such cross-level connections via conceptual analysis. So a more precise version of the argument looks as follows:

Transparent reductive explanation of 'water boils' – ***second pass***

Premise 1 (deducible from the base set):

H_2O is engaged in behaviour B at 212 °F at sea level.

Premise 2 (from *a priori* conceptual analysis):

Water is the stuff that plays the water role.

Premise 3 (deducible from the base set):

H_2O is the stuff that plays the water role.

Conclusion 1 (from Premise 2 and Premise 3):

H_2O is water.

Premise 4 (from *a priori* conceptual analysis):

Boiling is the process that plays the boiling role.

Premise 5 (deducible from the base set):

Molecular behaviour B is the process that plays the boiling role.

Conclusion 2 (from Premise 4 and Premise 5):

[3] Or, at best, a concept of phenomenological thermodynamics—a macro-level theory crucially different from the micro-level theory of statistical mechanics.

[4] In the context of the Block-Stalnaker argument, the same molecular counterpart of 'boiling'—called behaviour B here—was called 'molecular process MP'. Cf. §6.3.3.

Molecular behaviour B is boiling.

Conclusion 3 (from Premise 1, Conclusion 1, and Conclusion 2):

Water boils at 212 °F at sea level.

What happens here is this. The original Premise 1 of the transparent reductive explanation above claiming that 'H_2O boils at 212 °F at sea level' is problematic for at least two reasons. First, the predication it makes (i.e. H_2O *boils*) refers to a higher level phenomenon (i.e. *boiling*), and thus violates the very requirement proponents of transparent reductive explanation take really seriously: that the explanatory premises of the deductive argument should not refer to phenomena belonging to the target level. Second, in accordance with the structure of transparent reductive explanation, Premise 1 should be deducible from base level facts and theories ('chemical theory of H_2O' or statistical mechanics). However, since the premise in question refers to some target level activity, it cannot be deduced solely from the base level. The very gap lurking behind and jeopardising every reductive explanatory effort (of the heterogeneous sort) pops its head up here as well.

At a first pass, 'water boils' cannot simply be deduced from a molecular level theory, because molecular level theories do not use the term 'water'. For a deductive argument to go through, one needs to connect 'water' to those terms which *are* used at the molecular level (e.g. H_2O). This is recognised by the original transparent reductive argument, hence its amendment with the original 'Premise 2 – Premise 3 – Conclusion 1' triad generating the required connection between the levels of the explanandum and the explanans by identifying water with H_2O.

Now, at second pass, we realise that molecular level theories do not use the term 'boiling' either. So we need an extra explanatory 'triad', a new 'Premise N – Premise N+1 – Conclusion K' amendment generating a new connection between the levels of the explanandum and the explanans, this time by identifying boiling with its counterpart described in molecular terms. This counterpart is ‘molecular behaviour B’ in the above argument.

Note how difficult it is to clear the way for a proper reductive explanation. Even our second pass argument will *not* go through: its Premise 1 (‘H_2O is engaged in behaviour B at 212 °F at sea level’) is still not deducible from the base facts and theories, since the term ‘212 °F’ refers to a particular value of temperature, and ‘temperature’, just like ‘water’ and ‘boiling’, is not part of the base level vocabulary. What can be deduced from the base level theory instead, is a claim about ‘mean molecular kinetic energy’. That is, one needs another explanatory triad connecting temperature to mean molecular kinetic energy. And even more, since the term ‘sea level’ stands for a certain amount of atmospheric pressure (1 bar)—something which is also inaccessible within the base level vocabulary.[5] So the most promising candidate for a *proper* (i.e. not trivially unsound) transparent reductive explanation of ‘water boils at 212 °F at sea level’ looks something like this:

> **Transparent reductive explanation of ‘water boils’ – *third pass***
>
> **Premise 1** (deducible from the base set):
>
> *H_2O is engaged in behaviour B at mean molecular kinetic energy MKE when the average force exerted by molecular collisions at a unit surface area is F.*
>
> **Premise 2** (from *a priori* conceptual analysis):
>
> *Water is the stuff that plays the water role.*
>
> **Premise 3** (deducible from the base set):
>
> *H_2O is the stuff that plays the water role.*
>
> **Conclusion 1** (from Premise 2 and Premise 3):
>
> *H_2O is water.*
>
> **Premise 4** (from *a priori* conceptual analysis):
>
> *Boiling is the process that plays the boiling role.*

[5] Pressure in the sense of ‘atmospheric pressure’ is a term of phenomenological thermodynamics and is defined via the volume and the temperature of a body of gas. The term ‘pressure’ as utilised by statistical mechanics, on the other hand, is defined within an entirely different vocabulary, via the amount of force exerted by molecular collisions on a surface area.

Premise 5 (deducible from the base set):

Molecular behaviour B is the process that plays the boiling role.

Conclusion 2 (from Premise 4 and Premise 5):

Molecular behaviour B is boiling.

Premise 6 (from *a priori* conceptual analysis):

'212 °F' is the stuff that plays the 'temperature 212 °F' role.

Premise 7 (deducible from the base set):

'Mean molecular kinetic energy MKE' is the stuff that plays the 'temperature 212 °F' role.

Conclusion 3 (from Premise 6 and Premise 7):

Mean molecular kinetic energy MKE is 212 °F.

Premise 8 (from *a priori* conceptual analysis):

'1 bar pressure' is the stuff that plays the 'pressure$_{thermodynamics}$ 1 bar' role.

Premise 9 (deducible from the base set):

'Average force F exerted by molecular collisions at a unit surface area' is the stuff that plays the 'pressure$_{thermodynamics}$ 1 bar' role.

Conclusion 4 (from Premise 8 and Premise 9):

Average force F exerted by molecular collisions at a unit surface area is 1 bar pressure.

Conclusion 5 (from Premise 1, Conclusion 1, Conclusion 2, Conclusion 3, Conclusion 4 and the target level observation that the pressure at sea level is 1 bar):

Water boils at 212 °F at sea level.[6]

As complicated as it might be, proponents of the *a priori* passage view are perfectly happy with this argument. The 'triads' 'Premise 2 – Premise 3 – Conclusion 1', 'Premise 4 – Premise 5 – Conclusion 2', 'Premise 6 – Premise 7 – Conclusion 3' and 'Premise 8 – Premise 9 – Conclusion 4' provide the *a priori* passages from H_2O to

[6] Of course, putting things this way is still inaccurate. Throughout the dissertation I greatly simplify the base level expressions. So, for example, it is not H_2O that plays the water role, and hence is identical with water, but rather a 'certain physical state of an aggregate or collections of H_2O molecules'.

water, from specific molecular behaviour to boiling, from mean molecular kinetic energy to temperature, and from force exerted on a surface to thermodynamical pressure respectively. With all these in place, the argument runs through smoothly, and so the particular water fact in question becomes transparently reductively explainable in terms of H_2O facts—or does it?

Note that with these purported *a priori* passages come the additional premises Premise 3, Premise 5, Premise 7 and Premise 9, all implying that a certain statement is deducible from the facts and theories of the base level. The content of these statements is crucial for the fate of the transparent model of reductive explanation. The devil is in the details of how the 'water role', 'boiling role', 'temperature role', etc. are spelled out. Here, different versions of the transparent model of reductive explanation differ significantly. In this section I am going to concentrate on what I call the 'standard model' of transparent reductive explanation featuring, for example, in the early writings of Chalmers (1996) and Levine (1983, 1993). In the next section (§7.1.2) I will compare this version with Kim's functional reduction, and finally, in §7.1.3, I will investigate how Chalmers and Jackson (2001) try to overcome the difficulties detailed below.

According to the 'standard model' of transparent reductive explanation, 'water role', 'boiling role', 'temperature role' etc. are shorthands for a list of the reference fixers of 'water', 'boiling', 'temperature', etc. respectively (cf. §6.3.1). For example, the reference fixers of 'water', as Jackson argues, include: "being potable, odourless, falling from the sky, being the stuff that makes up various bodies of liquid of our acquaintance" (Jackson, 2003, p. 86). So Premise 3 above can be spelled out in this way: 'H2O is the stuff that is potable, odourless, falling from the sky, etc'.

Now note the crucial fact: 'potable', 'odourless', and similar terms like 'transparent', 'colourless' etc., which typically feature in the different examples of how the 'water role' (or 'watery stuff') can be spelled out are—in a fundamental sense—just like ‘water’. They all belong to the same ballpark of concepts, so to speak, i.e. they belong to the vocabulary utilised at the target level. That is, the terms ‘potable’, ‘odourless’, etc. refer to target level phenomena. Moreover, these terms cannot be found in the vocabulary of the base level. The chemical theory of H_2O, as Levine would put it, or statistical mechanics, as others prefer to call it—i.e. the base level theory,—tells us nothing about being ‘potable’ or ‘odourless’. Just as there is nothing described as ‘water’ at the base level, there is nothing ‘odourless’ at the base level, and just as H_2O does not boil, H_2O is not ‘transparent’ either. That is, it is impossible to deduce only from the base level facts and theories the statement that ‘H_2O is odourless’, or that ‘H_2O is transparent’.

But wait, wasn’t the sole purpose of introducing Premise 3 (along with Premise 2 and hence Conclusion 1) to make the move from the base level to the target level possible? Remember, the very task at hand is to account for higher level phenomena in terms of lower level phenomena. The difficulty with this task is that in almost all interesting cases[7] the descriptions available at the higher target level utilise a different vocabulary than those available at the lower base level. That is, it is not trivial to decide whether it is the same fact that the two different descriptions of the target and the base level talk about—let alone identifying which description of the base level corresponds to which description of the target level. This is the so-called ‘gap’ between the target and the base level (cf. §6.3.1), which needs to be bridged by any attempt of reductive explanation.

The standard model of transparent reductive explanation claims that it is able to bridge this gap by conceptual analysis, which amounts to amending the original deductive argument with ‘Premise N – Premise N+1 – Conclusion K’ triads consisting of an *a priori* conceptual analysis of a higher level concept in terms of its reference fixers, and an *a priori* deduction solely from the facts and theories available at the base level of the claim that the particular role defined by the

[7] Nagel calls these *heterogeneous* cases (E. Nagel, 1961).

reference fixers is filled by a certain base level entity or process. As we have just seen, however, the standard model of transparent reductive explanation runs into difficulties regarding the latter *a priori* deduction: the statements that ought to be deduced deploy terms which belong to the target vocabulary and thus cannot be deduced solely from the base.

That is, the purportedly deducible premise of the amendment is, in fact, not deducible. There is a gap between the base level facts and theories, on the one hand, and the statement to be deduced from them due to the fact that they utilise different vocabularies. Without bridging this gap, the deductive argument as a whole cannot go through. But the only tool the standard model of reductive explanation can rely on to bridge this gap is the same conceptual analysis based method that has been deployed in the first instance. That is, for the argument to go through, the standard model of reductive explanation needs to amend the argument with yet again a new 'Premise N – Premise N+1 – Conclusion K' type triad, where the new Premise N is an *a priori* conceptual analysis of the target level concept featuring in the Premise N +1 of the previous iteration in terms of *its* reference fixers, and the new Premise N+1 is an *a priori* deduction solely from the facts and theories available at the base level of the claim that the particular role defined by the reference fixers of the target level concept deployed by the previous iteration of Premise N+1 is filled by a certain base level entity. This whole process, of course, leads to a vicious circle.

To illustrate the problem with a concrete example, consider how the water case escalates into a vicious circle. As we have seen, to get from H_2O to water, one needs to be able to get e.g. from H_2O facts and theories to potability. However, in order to do so, first one needs to analyse 'potable' in terms of its reference fixers. Such an analysis could be provided along the following line: 'potable is the stuff that can be digested, is fluid, non-toxic, etc.'. As a next step, one also needs to be able to get from H_2O facts and theories to digestibility, fluidity, etc. Now since digestibility and fluidity are water-level phenomena, and the related terms do not feature in the vocabulary of H_2O level facts and theories, it is impossible to deduce digestibility-involving claims from the base level facts and theories, unless one provides a conceptual analysis of digestibility in terms of e.g. consumption and secretion, and a corresponding deduction from the base level facts and theories how a particular base level entity fills the role defined by consumption and secretion, etc. Which, in turn—since 'consumption' and 'secretion' are not parts of the base level vocabulary—is, of course impossible unless yet another layer of conceptual analysis comes to the rescue.

Within the framework of the standard model of reductive explanation, this problem seems to be in principle unsolvable. The reason for this is the following: for Premise N to be *a priori* available via conceptual analysis, it needs to rely on such reference fixers which belong to the same ballpark of concepts as the very concept under analysis. The problem of the different vocabularies is so pressing exactly because there are no cross-vocabulary conceptual connections between concepts of the target domain and concepts of the base domain. Any given macro-concept can be analysed *a priori* only in terms of other macro-concepts. If it were otherwise, figuring out the chemical underpinnings of biology, or doing quantum chemistry would be a matter of armchair conceptual analysis.

That is, the transparent version of reductive explanation in general faces a dilemma: either Premise N is *a priori* available via conceptual analysis, in which case, the reference fixing roles get determined by target level concepts, and thus Premise N+1 becomes unavailable at the base level, or we try to make Premise N+1 available at the base level by determining the reference fixing role of the problematic target level concept in terms of base level concepts, in which case Premise N becomes unavailable. As we have seen in this section, the so-called standard version of transparent reductive explanation is on the first horn of this dilemma. In contrast with this, Kim's functional model of reductive explanation is on the second horn. I explain why in the next section.

7.1.2 Kim’s functional reduction

On the face of it, Jaegwon Kim’s functional model of reduction is quite similar to what I have been calling the ‘standard model’ of reductive explanation. To bridge the gap between the levels of the explanandum and the explanans (or as Kim (2005) refers to it, to secure the explanatory ascent) the standard model introduces a ‘Premise N – Premise N+1’ pair (bringing along Conclusion K), which first connects the target level concept to a specific role, and then determines what base level entity fills that particular role. As we have seen in §6.2.3, Kim’s model has the same structure. The first two steps of the three-step process of functional reduction look as follows:

> “Step 1 [Functionalization of the target theory]
> Property M to be reduced is given a *functional definition* of the following form:
> Having M = (def.) having some property or other P (in the reduction base domain) such that P performs causal task C. […]
>
> Step 2 [Identification of the realizers of M]
> Find the properties (or mechanisms) in the reduction base that perform the causal task C.”
> (Kim, 2005, pp. 101-102, original emphasis)

Step 1 (just like Premise N of the standard model) connects the target phenomenon (property M, in this quote) to a particular role. This role is explicitly a functional, or causal role in Kim’s case. So what step 1 does is re-defining the target phenomenon in terms of this functional role; hence the name ‘functionalisation’. Step 2, then, (again, just like Premise N+1 within the standard model) identifies that particular base level entity which fills the role definitive of the target phenomenon. These two steps together create the required connection—the bridge—between the target and the base levels. Given these similarities, it is no surprise that Kim (2005) actually spends a considerable amount of time drawing analogies between his model and other descriptions of reductive explanation put forward by Chalmers (1996) and Levine (1983, 1993).

However, there is a fundamental difference between the ‘standard model’ of transparent reductive explanation and the view advocated by Kim. If we take a closer look at Kim’s step 1, what we find is this: “Having M = (def.) having some property or other P (in the reduction base domain) such that P performs causal task C” (Kim, 2005, p. 101). This definition ties the target phenomenon to a causal task performed by an entity *of the base level*. That is, the causal task in terms of which Kim’s functionalisation re-defines the target level phenomenon belongs to the base level. This is in stark contrast with how Premise N of the standard model connects the target phenomenon to a list of reference fixers described in the vocabulary of the target level.

At a first pass, this move might seem promising: it seems to open up the possibility of solving the problem of the standard model and rescuing transparent reductive explanation from the vicious circle. Remember, the original problem with the standard model was that providing an analysis of the target concept in terms of other target level concepts does not help at all in bridging the gap between the levels of the target and the base. Now if Kim’s functional model is able to re-define the target level phenomenon in terms of a base level causal role, then that might do the trick.

Unfortunately (for the proponents of transparent reductive explanation), it isn’t. Note that the process of functionalisation is an *intra*-theoretic process. As we have already seen, in typical cases of scientific reduction (remember, Kim proposes his account as a model of scientific reductions) there are different theories in operation at the target and the base level. And it is the theory in operation at a given level which determines the kinds of causal relationships that can obtain between the entities of that level. That is, when Kim re-construes the target phenomenon as defined by causal relations to entities in the reduction base, he glosses over the differences between the interactions of the target entities and the interactions of the base entities as determined by the different target and base level theories. One cannot functionalise a target entity relating it to the entities of the base theory, since one cannot functionalise an entity of a given theory relating it to entities of another theory. What is possible is to functionalise the target entity and separately to functionalise the base entities; that is, to (re-)define the target entity by its causal relations to other entities of the *target theory*, and to outline a similar causal-relational network of the entities in the base theory. So any statement expressing the causal role of the target entity that *can* be formulated must rely only on entities and interactions of the target theory, and similarly any statement expressing the functional relationships in the base theory must rely only on entities and interactions of the base theory.

That is, instead of a single functional role C characteristic of both the target property M and the corresponding base property P, as Kim in the quote above assumes, in fact, there are two different functional tasks (at least in the interesting heterogeneous cases, where the vocabularies of the two theories significantly differ): C_T of M in the target T-domain, and C_B of P in the base B-domain. M fulfils the causal roles described by $C_T = \{c_{T1},\ldots, c_{Ti}, e_{T1}, e_{Ti}, \ldots\}$ in a way that $\{c_{T1},\ldots, c_{Ti}\}$ cause M to be instantiated and M causes $\{e_{T1},\ldots, e_{Ti}\}$ to be instantiated, and so does P with $C_B = \{c_{B1},\ldots, c_{Bj}, e_{B1},\ldots, e_{Bj}\}$. Consider, for example, M = heat and P = mean molecular kinetic energy. C_T connects heat with pressure, volume etc. C_B connects mean molecular kinetic energy with velocity, mass, etc.[8] These are different sets of theoretical entities described by different theories.

Let's pause here for a moment. Kim tries to connect the base and the target levels by the first two steps of his functional model of reduction. In the first step, he (re-)defines the properties picked out by the theoretical expressions of the target theory in terms of the causal roles they play. In the second step, he tries to find the relevant causal roles amongst the causal roles of the properties picked out by the theoretical expressions of the reducing base theory. However, in order to functionalise a property picked out by a theoretical term of the target theory one needs to track down

[8] The target level theory determines, for example, that 'if the temperature of a gas is increasing then, given a fixed volume, the pressure of the gas will increase too', whereas the base level theory determines that 'if the mean kinetic energy of a bunch of molecules in a tank increasing then the number of the impacts on the wall of the tank will increase too'.

the causal connections this property has—and any property *as picked out by a term of the target theory* has causal connections only with other properties picked out by other terms of the target theory. That is, the causal roles of the properties as picked out by the expressions of the target theory are determined by the target theory itself. In the same way, the causal roles of the properties picked out by the terms of the base theory are determined by the base theory itself. C_T and C_B are thus causal roles defined by different theories: the target theory and the base theory respectively. $\{c_{T1}, \ldots, c_{Ti}, e_{T1}, \ldots, e_{Ti}\}$ are determined within the target theory, whereas $\{c_{B1}, \ldots, c_{Bj}, e_{B1}, \ldots, e_{Bj}\}$ are determined within the base theory.

What Kim has to confront here is the second horn of the dilemma discussed at the end of the previous section: if one tries to make the claim connecting a certain target level entity to a particular causal role deducible at the base level by determining the causal role in terms of base level concepts, then there will be no connection between this specific causal role and the very target level entity one wanted to tie to the base level in the first place.

7.1.3 The Chalmers-Jackson proposal

One of the major criticisms Block and Stalnaker (1999) have formulated against transparent reductive explanation claims that the kind of explicit conceptual analysis the standard model relies on is in general unavailable.[9] Chalmers and Jackson (2001) in their response defend the *a priori* passage view by arguing that it does not need to rely on explicit conceptual analysis. At first sight, their proposal is a good candidate for rescuing transparent reductive explanation from the dilemma both the standard model and Kim's model have to face with. In order to evaluate whether Chalmers and Jackson succeed in solving the problem discussed in the previous sections, let's consider their proposal in detail.

[9] Cf. Footnote 21 in §6.3.2.

Behind the proposal, there is a fundamental analogy Chalmers and Jackson rely on. They invite us to consider Gettier's second case (cf. Gettier, 1963), and argue, on the basis of that example that *a priori* entailment is available even without explicit conceptual analysis. They formulate this argument in the following way:

> "Let G be the conjunction of the statements in the following passage: 'Smith believes with justification that Jones owns a Ford. Smith initially has no beliefs about Brown's whereabouts. Smith forms a belief that Jones owns a Ford or Brown is in Barcelona, based solely on a valid inference from his belief that Jones owns a Ford. Jones does not own a Ford, but as it happens, Brown is in Barcelona'. Let K be the statement '[Smith] does not know that Jones owns a Ford or Brown is in Barcelona'. It is plausible that [G entails K] is a priori. But it is also plausible that there is no explicit analysis of the concept of knowledge using the terms involved in G. If so, a priori entailment does not require explicit analyses of the terms in the consequent using the terms in the antecedent. It is also somewhat plausible that there is no explicit analysis of the concept of knowledge at all. If so, a priori entailment does not require explicit analyses of the terms in the consequent." (Chalmers & Jackson, 2001, pp. 320-321)

In this example G is a description of Gettier's second case. The description utilises terms like 'belief', 'justification', 'valid inference', etc., *but not* 'knowledge'. K is a statement involving the term 'knowledge'. Chalmers and Jackson argue that "it is plausible that G *a priori* entails K". However, at the same time, as Gettier shows us, the concept of knowledge cannot be analysed in terms of justified true belief, and moreover, as Chalmers and Jackson further argues, it is also plausible that there is no other "explicit analysis of the concept of knowledge using the terms involved in G". If this is true, then *a priori* entailment is possible without an explicit conceptual analysis of the target concept in terms of the concepts utilised in the base. If the further claim, according to which "it is also somewhat plausible that there is no explicit analysis of the concept of knowledge at all" is also true, then *a priori* entailment is possible even without any kind of an explicit conceptual analysis of the target concept (not even in terms of other concepts of the target domain).

Compare now this to what we have learned so far of transparent reductive explanation. According to Kim's model, the target phenomenon must be re-defined in terms of base level causal roles. Since Kim thinks of this kind of functional analysis as a version of conceptual analysis, Kim's model requires an explicit conceptual analysis of the target concept in terms of concepts utilised at the base level. On the other hand, the standard model of transparent reductive explanation is committed to an explicit conceptual analysis of the target concept in terms of other concepts of the target level (cf. the utilisation of a list of reference fixers). So the Chalmers-Jackson proposal, which claims that *a priori* entailment is possible without any kind of explicit conceptual analysis, is a true alternative to both Kim's model and the standard model.

Note, however, how radical this proposal is, compared to what either the standard model or Kim's approach says about how transparent reductive explanation naturally proceeds. As we have seen, the major trouble every reductive attempt needs to overcome is bridging the gap between the target and the base domains. Conceptual analysis serves this very purpose in transparent reductive explanation. According to the standard model, the target and the base is connected by the application of the conceptual analysis of the target level concept (originally formulated at the target level utilising other target level concepts) at the level of the base domain. Although Kim's approach disagrees, and analyses the target level concept directly in terms of base level concepts, it is only a minor difference regarding how conceptual analysis should be performed. The two versions of transparent reductive explanation do not dispute the role of conceptual analysis—they both agree that without conceptual analysis the target and the base domains cannot be connected.

Contrary to all this, Chalmers and Jackson's proposal claims that there is no need for an explicit bridge between the target and the base domains. They argue that it is a fundamental fact about concepts: once one masters a concept, one is able to decide whether that specific concept can be applied in a situation originally described by a different vocabulary. As Chalmers and Jackson put it:

> "[It is a] general feature of our concepts. If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension. For example, if a subject possesses the concept 'water', then sufficient information about the distribution, behavior, and appearance of clusters of H_2O molecules enables the subject to know that water is H_2O, to know where water is and is not, and so on [...] a 'water'-free description of the world can enable one to identify the referent of 'water', and a 'knowledge'-free description of the world can enable one to decide whether a given belief is an instance of knowledge. In these cases, we can say that nontrivially sufficient information enables identification of a concept's extension." (Chalmers & Jackson, 2001, p. 323)

That is, Chalmers and Jackson claims that if we master the target level concept, and are given a base level description rich enough, then we will be able to pick out that bit of the base level description which co-refers with the target level concept.[10] This allows us to transfer the explanatory power of the base level to the target level without further ado. Consequently, transparent reductive explanation, according to Chalmers and Jackson, looks as follows:

> **Transparent reductive explanation of 'water boils' – *C&J interpretation***
>
> **Premise 1** (deducible from the base set):
>
> *H_2O is engaged in behaviour B at mean molecular kinetic energy MKE when the average force exerted by molecular collisions at a unit surface area is F.*
>
> **Conclusion 1** (from mastering 'water', and a rich-enough description of the base level):
>
> *H_2O is water.*

[10] The view that one does not need to rely on explicit conceptual analysis, i.e. that for any concept subjects possessing the concept are able to identify the extension of the concept in a sufficiently rich description is called *ascriptivism*. As Diaz-Leon summarises, ascriptivism is the view that "[f]or any concept *C*, there is an application condition like this: (AC): 'If *x* is *F*, then *x* falls under *C*'." (Diaz-Leon, 2011, p. 102) For the denial of this claim, i.e. for a defence of *non-ascriptivism*, see, for example, Levine (2001).

Conclusion 2 (from mastering ‘boiling’, and a rich-enough description of the base level):

Molecular behaviour B is boiling.

Conclusion 3 (from mastering ‘temperature’, ‘fahrenheit’, etc., and a rich-enough description of the base level):

Mean molecular kinetic energy MKE is 212 °F.

Conclusion 4 (from mastering ‘pressure’, ‘bar’, etc., and a rich-enough description of the base level):

Average force F exerted by molecular collisions at a unit surface area is 1 bar pressure.

Conclusion 5 (from Premise 1, Conclusion 1, Conclusion 2, Conclusion 3, Conclusion 4 and the target level observation that the pressure at sea level is 1 bar):

Water boils at 212 °F at sea level.

In this model of transparent reductive explanation, no extra premises are added to the argument creating a connection between the target and the base level. Any subject mastering the target level concepts and having “unimpaired rational processes” is able to conclude on the identities; i.e. is able to read off from the base level description, which base level expression corresponds to which target level concept.

If this model of transparent reductive explanation is correct, then Chalmers and Jackson do not need to worry about the dilemma introduced above. As we have seen, the dilemma itself is a consequence of the need for a bridge connecting the target level concept and the base level description. If, however, no such bridge is necessary, then, of course, there is no dilemma either. So, it looks as though the version of transparent reductive explanation Chalmers and Jackson propose outperformed both the standard model and Kim’s approach as well.

In what follows, I will argue that this appeal of Chalmers and Jackson's proposal is only superficial. Once one analyses the notion of concept-mastery they heavily rely on, it turns out that Chalmers and Jackson are not able to avoid the very dilemma that has already proven fatal for the other two versions of transparent reductive explanation.

First of all, consider the following passage from the quote above:

> "For example, if a subject possesses the concept 'water', then sufficient information about the distribution, behavior, and appearance of clusters of H_2O molecules enables the subject to know that water is H_2O, to know where water is and is not, and so on." (Chalmers & Jackson, 2001, p. 323)

What happens here is that the subject, by possessing the concept 'water', becomes able to pick out H_2O as the base level term co-referring with 'water'. According to Chalmers and Jackson, by mastering a target level concept, one becomes able to identify the corresponding base level expression. This is very similar to how a conceptual analysis of a target level concept allows proponents of the standard model to pinpoint the corresponding base level term. That is, in Chalmers and Jackson's framework concept-mastery plays the role explicit conceptual analysis plays in the standard model of transparent reductive explanation—i.e. this is the tool that connects the target and the base levels.

Now as we have seen, the idea of connecting the target and the base levels via conceptual analysis faces a dilemma. Either the conceptual analysis of the target level concept is *a priori* available (cf. standard model), in which case, the reference fixers get determined by target level concepts, and thus these reference fixers cannot be straightforwardly applied at the base level, or one tries to determine the reference fixers of the problematic target level concept in terms of base level concepts (cf. Kim's approach), in which case the conceptual analysis (functional re-definition) of the target concept becomes unavailable. Can the notion of concept-mastery overcome this obstacle?

It is hard to see how concept-mastery could ever get beyond what is depicted by the list of reference fixers of a concept within the standard model. True, the notion of concept-mastery does not presuppose an explicit list of reference fixers, i.e. it is compatible with cases where the subject is able to apply the concept correctly without checking the actual situation against an explicit list of application conditions (that is, the subject 'just knows' when to apply the given concept). This makes Chalmers and Jackson's proposal a valid answer to Block and Stalnaker's initial criticism.[11] However, the problem the afore-mentioned dilemma poses is quite different from Block and Stalnaker's challenge. The question is not whether giving an explicit list of reference fixers is possible, but whether it is possible to match reference fixers as described in one vocabulary with concepts belonging to another vocabulary.

To see how this latter question is related to Chalmers and Jackson's proposal, consider what possessing or mastering a concept amounts to. Of course, the way one thinks about concepts in general greatly affects what one thinks about this particular issue. However, under contemporary theories of concepts, categorisation, i.e. the application of a concept is based on certain core features—application conditions which have to be fulfilled in order to trigger the actual use of the concept (cf. Prinz, 2002). These core features guiding the application of a given concept are tied to the characteristics of e.g. perceptual prototypes, and the most typical scenarios in which the concept is applied. Note that these characteristics and scenarios, when turned explicit, are described with the aid of the very same vocabulary to which the original concept belongs. For example, the concept 'water' is most typically applied when someone sees something which is transparent, liquid, odourless, tasteless, etc. and is placed in a cup, flows from the tap, or faces us on the beach. That is, when someone is given a sufficiently detailed description of the distribution, behaviour, and

[11] Claiming that an explicit conceptual analysis in not always available—cf. Footnote 21 in §6.3.2.

appearance of such a liquid, one is able to know where water is and is not, etc. (cf. the quote above).

The problem is that the subject is in an entirely different situation if she is given a sufficiently detailed description of the distribution, behaviour, and appearance of *H_2O molecules*. Such a description would talk about the different characteristics (e.g. position, impetus, energy, etc.) of molecules, and say nothing about being transparent, odourless, etc. This is a point which has already been emphasised in the previous sections—the very source of the dilemma poisoning transparent reductive explanation: in the typical, and most interesting cases of scientific reductions and purported reductive explanations, the base and the target level descriptions utilise entirely different vocabularies. In the present context, the result is that the description of H_2O molecules, no matter how detailed it is, will be very unfamiliar for the subject, will clash with the application conditions of the concept 'water', and thus will not trigger the use of the concept. In other words, Chalmers and Jackson misdescribes the situation. Contrary to what they claim, "if a subject possesses the concept 'water', then" even available "information about the distribution, behavior, and appearance of clusters of H_2O molecules" *will not* enable "the subject to know that water is H_2O, to know where water is and is not, and so on" (Chalmers & Jackson, 2001, p. 323).

However, Chalmers and Jackson might want to dig their heels in here and argue that the theory of concepts evoked above is not the right one. For their standards, the set of application conditions of a concept is not exhausted by features of the most typical scenarios, but contains all the features of all the possible scenarios in which the concept can be applied. That is, only those subjects qualify as possessors of a given concept who are able to identify the extension of a given concept if presented with a sufficiently rich description—even if the description is formulated entirely with the vocabulary of a lower level. According to this view, then, those unable to identify the extension of, say, the concept 'water' if presented with a full microphysical description of the world, in fact, do *not* possess or master the concept in question.

Note, however, how unintuitive this view is. My grandma, for example, has never ever heard of molecules, let alone microphysics. I am sure that if she was provided with a description of the behaviour of a collection of H_2O molecules, then she would be, well, quite confused. She would not be able to make any sense of it. Nonetheless, she has lived a whole lifetime during which she has used the concept 'water' many times. She knows what to do when I ask for a glass of water. She knows how to use it when making a soup. She even understands the word 'billabong', a word she has never ever heard before, if I explain it using the word 'water' along with others like 'still', 'river', 'cut off', etc. My grandma masters the concept 'water'. By any *intuitive* standard, she possesses this concept.

Esa Diaz-Leon (2011) calls the view attributed to Chalmers and Jackson above, namely that lower level application conditions are part of a concept's possession conditions *reductive ascriptivism*.[12] Diaz-Leon argues in length that reductive ascriptivism is unintuitive. She uses the example of the concept 'square', and a corresponding application condition "If x_1, x_2 ... x_n instantiate properties F_1, F_2 ... F_n, then r is square" (Diaz-Leon, 2011, p. 107), where x_1, x_2 ... x_n are microphysical entities and F_1, F_2 ... F_n are microphysical properties. She asks us to imagine Tina, a subject who does not know this application condition. Nonetheless, Tina is able to entertain thoughts involving the concept. As Diaz-Leon puts it:

> "It seems perfectly compatible with lacking such a conditional ability that she could in effect use the word 'square' very much as other subjects who have the conditional ability would. She can have conversations with those subjects, in which they apply the word 'square' to the same objects. If Tina did not possess

[12] As she puts it: "[Reductive ascriptivism is the view that] for any concept *C* at level *n*, there is an application conditional like this: (AC): 'If *x* is *F*, then *x* falls under *C*', where *F* is described using concepts from a lower level *m*" (Diaz-Leon, 2011, p. 109). Cf. Footnote 10 above for Diaz-Leon's definition of ascriptivism.

> the concept SQUARE, then how can she understand what the others are saying? How can she communicate with them?" (Diaz-Leon, 2011, p. 110)

From this Diaz-Leon concludes that Tina, in fact, possesses the concept 'square'. In drawing this conclusion, Diaz-Leon explicitly relies on a principle proposed by Williamson, according to which "[i]f one understands the word 'C', one has the concept C" (Williamson, 2003, p. 290; cited by Diaz-Leon, 2011, p. 111). If, however, subjects who do not know the microphysical application conditions of a given concept nevertheless possess the concept, then reductive ascriptivism cannot be true. But then, the only way ascriptivism itself (i.e. the view that concept possession does not presuppose explicit definitions) can be true is if it is restricted to application conditions utilising same-level concepts.[13]

This then leaves us with the problem already discussed. Since in the typical heterogeneous cases of scientific reductions and reductive explanations the base and the target level descriptions utilise entirely different vocabularies, reductive explanations need to bridge this gap between the base and the target levels. Chalmers and Jackson's proposal—emphasising ascriptivism over non-ascriptivism—is no better candidate for bridging this gap than either the 'standard' version or the Kimian version of transparent reductive explanation. Chalmers and Jackson's version relies on concept ascription based on application conditions, but as we have seen, since target level concepts cannot be ascribed on the basis of base level application conditions, the Chalmers-Jackson proposal leaves the base and the target levels unbridged.[14]

[13] Diaz-Leon calls this version of ascriptivism *non-reductive ascriptivism*: "For any concept *C* at level *n*, there is an application conditional like this: (AC): 'If *x* is *F*, then *x* falls under *C*', where feature *F* is described using only concepts at level *n*." (Diaz-Leon, 2011, p. 109)

[14] Diaz-Leon argues that the Chalmers-Jackson proposal is unable to overcome this obstacle even by restricting the possessors of a given concept to experts. Experts are "those who know what properties determine whether something falls under a certain concept, and therefore they could find out the extension of such a concept, given a sufficiently rich description of a scenario" (Diaz-Leon, 2011, p. 111). This move, Diaz-Leon argues, is problematic for at least two reasons. First, it would trivialise the notion of *a priori* knowledge by rendering empirical discoveries instances of *a priori* knowledge. Second, it would result a technical notion of apriority which would, in turn, make Chalmers and Jackson's further claim that phenomenal truths are not *a priori* knowable on the basis of physical truths implausible (cf. Diaz-Leon, 2011, pp. 111-115).

Finally, note that Chalmers and Jackson even fail to provide a single example supporting their claim. The core example they rely on, i.e. Gettier's second case, has no implications whatsoever regarding the issue at hand: the heterogeneous cases of reductive explanation with different base and target level vocabularies. Even if they are right in their analysis that one is able to apply correctly the term knowledge in a situation described in a 'knowledge-free' vocabulary, from this alone it does not follow that one is able to apply any concept in any situation described in an appropriate 'given-concept-free' vocabulary. The difficulty all attempts of reductive explanation have to face with, as we have seen it, stems from the fact that the target concept belongs to a different vocabulary than those terms which are utilised to form the base level description. In this respect, Gettier's second case is crucially different from the issue at hand. The terms 'knowledge', 'belief', 'justification', etc. belong to the same ballpark of concepts, i.e. to the same vocabulary. Even if we cannot come up with an explicit analysis of one of them in terms of the others, still, there are inherent links between them. For example, accidentally true beliefs do not constitute knowledge—this much we know, and it plays a part in the application conditions of 'knowledge'. This is the reason why we intuitively judge Gettier's second case as such in which the concept 'knowledge' is not applicable. That is, even if the concept of 'knowledge' cannot be explicitly analysed in terms of 'belief', 'justification', etc., there are conceptual links between these terms. No such links exist between a concept belonging to one vocabulary (e.g. 'water' of folk vocabulary, or 'temperature' of thermodynamics) and a set of other concepts belonging to a different vocabulary (e.g. 'H_2O' of chemistry, or 'mean molecular kinetic energy' of statistical mechanics).

In other words, not even the Chalmers-Jackson proposal is immune to the threat due to the fact that different vocabularies are in use at the target and the base level. On the contrary, just as the standard model and Kim's version of transparent reductive explanation, the Chalmers-Jackson proposal too has to face with, and ultimately is invalidated by, the dilemma: if, on the one hand, we want to ensure that the subject becomes able to apply the target concept, then we have to provide a description utilising concepts belonging to the same vocabulary as the target concept, in which case the description provided will not talk about the base level as such, whereas, on the other hand, if we provide a proper description of the base level utilising base level vocabulary, then the subject will not be able to apply the target concept. In either way, there will remain an unbridged gap between the target and the base domains, and hence the Chalmers-Jackson version of transparent reductive explanation will fail.

7.2 Reductive Explanation via 'Prior Identities'

As a last resort, proponents of transparent reductive explanation might want to argue that even if it is true that there is a gap between a microphysical description utilising the vocabulary of, say, a molecular theory and our folk concepts, this gap evaporates once one considers the close relationships between different scientific theories at adjacent levels. In other words, the arguments says, it is possible to get from a micro-description to a macro-description via consecutive steps, with smooth transitions between the different vocabularies. And once one arrives at the macro-level, one becomes able to straightforwardly apply the folk concepts in question. That is, it seems possible to generate all the identities needed to get from the base level to the target level in small steps. Chalmers and Jackson have something very similar in mind. As they formulate it:

> "[I]n a reductive explanation of a phenomenon such as water or life, we find that a low-level account of the physical processes involved will in principle imply and explain truths about the macroscopic structure, dynamics, behavior, and [...] appearance of relevant systems. And our concepts of 'water' or 'life' dictate that systems with appropriate sorts of structure, dynamics, behavior, and appearance automatically qualify as water or as alive (at least if they are appropriately situated in our environment, or are relevantly related to systems in our environment)." (Chalmers & Jackson, 2001, p. 351)

Of course, this passage in itself is more like a testimony than an argument. It simply states that a low level account of physical processes in principle implies and explains truths about macroscopic features. So this claim needs some serious support. Let's see how Chalmers and Jackson back it up. They say:

> "[A] macroscopic description of the world in the language of physics is implied by a microscopic description of the world in the language of physics. Such a thesis is extremely plausible: it is *not subject to any worries about translation between vocabularies*, and *involves only a change of scale*. The only worry might concern the status of bridging principles *within physical vocabulary*: for example, is it a priori that the mass of a complex system is the sum of the masses of its parts?[15] If there are any concerns here, however, they can be bypassed by stipulating that the relevant physical principles are built into [physics]." (Chalmers & Jackson, 2001, pp. 330-331, emphases added)

Now this is interesting. Chalmers and Jackson seem to presuppose that micro and macro descriptions are all formulated with one single vocabulary: the "language of physics". This presupposition goes against the whole body of contemporary philosophy of science, where it is one of the most fundamental insights that in order to understand a complex macro system in terms of its constituents, one needs to rely on a plurality of different scientific theories. This is because different levels of description utilise different scientific theories—each with very different vocabularies.

In order to see this, in the next section I turn towards the most significant contemporary theory of scientific explanation, the so-called *mechanistic model of explanation* (Machamer, et al., 2000). My aim in this next section is to establish the claim that moving up from a micro description to a macro description step by step through adjacent levels is not that trivial as proponents of transparent reductive explanation presume, and thus has to face with the very same problem which we

[15] Based on this example of the mass of a complex system, what Chalmers has in mind here when he uses the term 'bridging principle' is not bridge laws in the Nagelian sense (cf. §6.2.1), but rather what Beckermann (2000) calls 'rules of combination'—rules determining how certain properties (e.g. scalars, vectors, etc.) combine.

have already identified: the problem of bridging the gap between different vocabularies.

7.2.1 Mechanistic explanations and prior identities

The mechanistic approach identifies a phenomenon via identifying the tasks performed—i.e. the causal roles played—by the phenomenon as a whole. The very point of the mechanistic approach is to explain a phenomenon by understanding the mechanism responsible for the task performed by the phenomenon.

Understanding the relevant mechanism consists in decomposing the phenomenon into parts, specifying the properties and causal roles of the parts, and understanding the spatial and temporal organisation of the parts. The mechanistic approach employs two interdependent building blocks for characterising mechanisms: entities and activities. Entities are the parts composing the mechanism, their activities are in virtue of what they contribute to the working of the mechanism. Entities and activities are interdependent because "entities and a specific subset of their properties determine the activities in which they are able to engage" and "activities determine what types of entities are capable of being the basis of such acts" (Machamer, et al., 2000, p. 6). A mere aggregate of entities and their activities is not a mechanism, however. Entities and activities must be arranged into a specific spatial and temporal order otherwise they would not be able to perform a certain task together. That is, the mechanistic approach tries to account for the phenomenon (the explanandum) in terms of the organised activity of its constituent parts.

Consider Figure 1[16]. S is the phenomenon in question, which gets identified via the task it performs—its ψ-ing. The arrows indicate how S ψ-ing is connected to its context, i.e. to other entities at its level. According to the mechanistic approach, S ψ-ing can be accounted for in terms of the components X_1, X_2, X_3, X_4 and X_5, and their activities (Φ_1-ing, Φ_2-ing, Φ_3-ing, Φ_4-ing, and Φ_5-ing respectively) organised in an

[16] Figure 1 is based on Figure 2 in (Craver, 2001, p. 66) and Figure 1.1 in (Craver, 2007, p. 7).

appropriate way. These entities and activities organised in the appropriate way *constitute* the mechanism responsible for the ψ-ing of S.

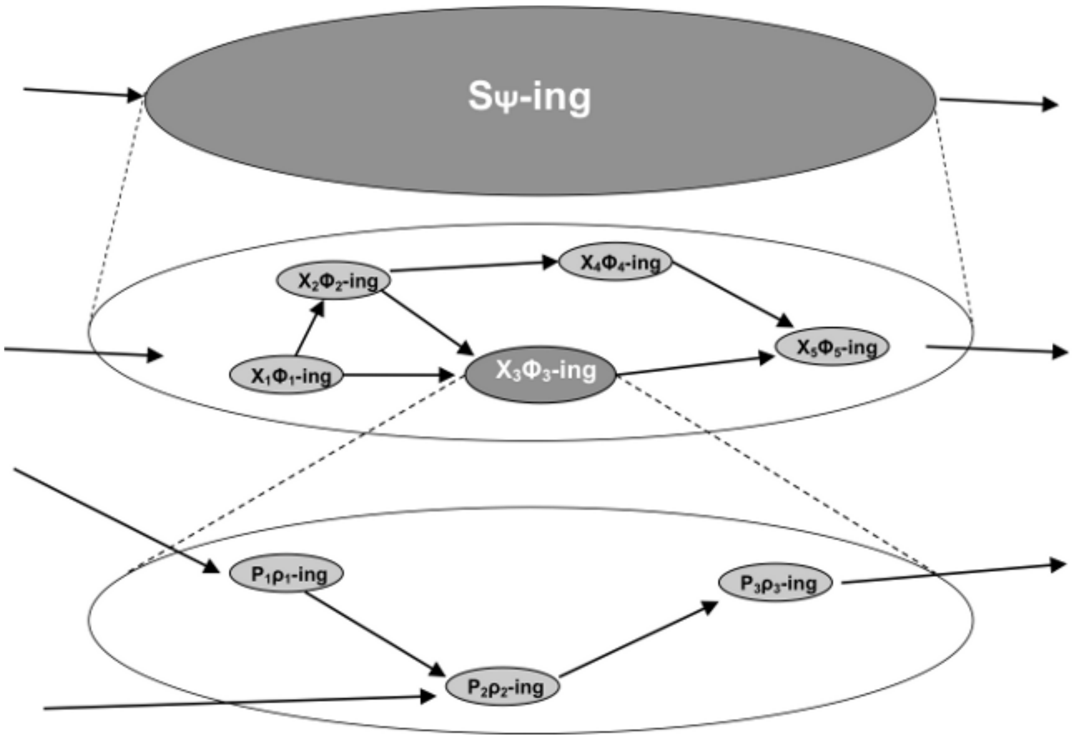

Fig. 1 The general scheme of mechanistic explanations.
S performs the task ψ-ing. On descending levels it is decomposed
into the organised activities of its constituent parts.

As it is shown on Figure 1, the mechanistic approach is a multilevel approach. Once one decomposed the ψ-ing of S into the lower level organised structure of entities (X_i-s) and their activities (Φ_i-ing) it is possible to apply the same methodology again in order to account for the Φ_i-ing of X_i-s in terms of a still lower level mechanism. So, for example, the Φ_3-ing of X_3 can be accounted for in terms of the organised ρ_i-ing of some P_i. P_1 ρ_1-ing, P_2 ρ_2-ing, and P_3 ρ_3-ing (organised in an appropriate way) together constitute the mechanism responsible for the Φ_3-ing of X_3.

In order to anchor the abstract characterisation introduced so far let's consider the following example: a person looking at Ishihara plates. What happens is that the person in question, when presented with an Ishihara plate and asked what she is seeing utters, say, 47. Ishihara plates contain a circle of dots of different sizes and colours. Some of the dots, coloured by similar shades, form numbers. Subjects with normal trichromat vision are able to group these dots together and recognise the number. So the phenomenon here is a person discriminating colours, which—in accordance with the mechanistic approach—gets identified via the task performed: once presented with an object stimulus the person makes a colour judgement. This level of description is called the level of the original phenomenon (L_P) on Figure 2.

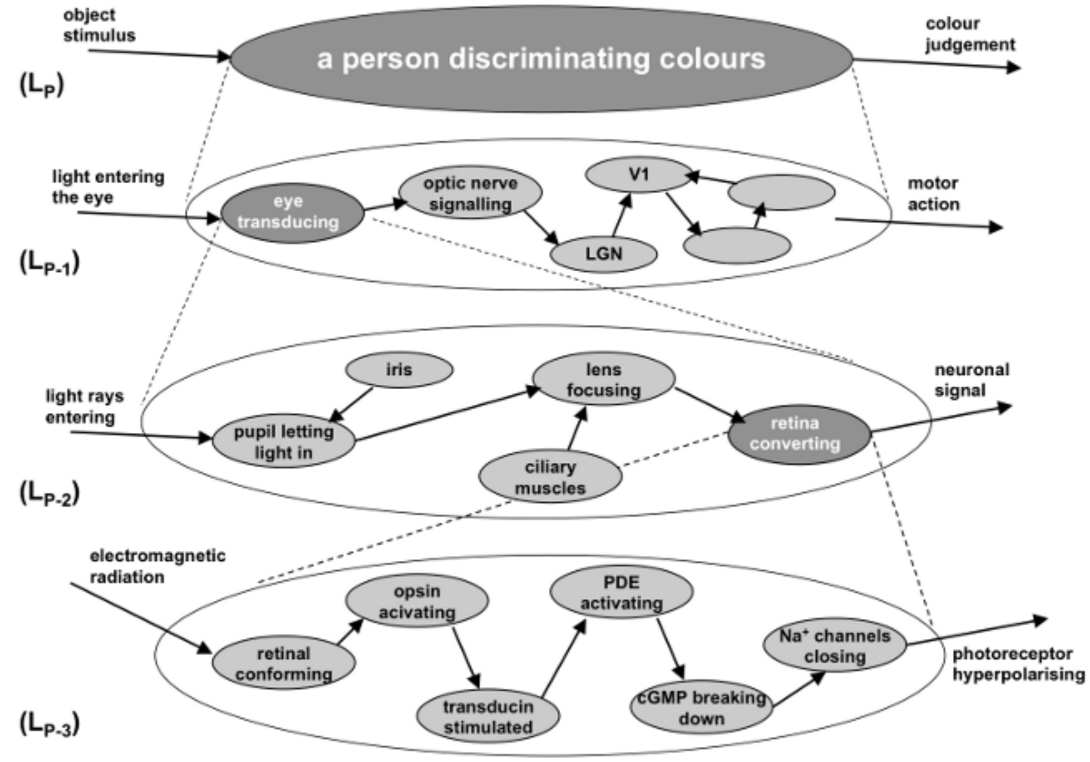

Fig. 2 Mechanisms of colour discrimination.

The mechanistic approach claims that explaining this phenomenon amounts to descending one level down and identifying a mechanism responsible for the task performed. This mechanism, roughly, consists in the light reflecting from the Ishihara plate and entering the eye, the eye transducing the energy of the light into the language of the nervous system, the optic nerve signalling and projecting to the lateral geniculate nucleus (LGN), the activation of the primary visual cortex (V1), then higher visual areas (especially V4) and other areas of the cortex, and finally the motor action of moving the mouth and the tongue (etc.) properly (sketched at the level L_{P-1} on Figure 2).

Each of these activities can be further analysed by descending one more level down (to level L_{P-2}), and identifying that mechanism which is responsible for the given activity at level L_{P-1}. So for example, the mechanism responsible for the activity of the eye is an organised activity of the parts of the eye. The pupil allows the light to enter the eye. The two muscles of the iris can vary the size of the pupil thus affecting the amount of light entering the eye. The lens inverts and focuses the light onto the retina. The ciliary muscles adjust the shape of the lens thus affecting its focus point. The retina converts the light into electrochemical signals. (Cf. level L_{P-2} on Figure 2.)

One need not stop here. Due to the advance of neurochemistry and molecular biology we are able to descend one more level down and identify the mechanism responsible for the activity of photoreceptors in the retina.[17] Electromagnetic radiation is absorbed by a photopigment in the membrane of the disks in the outer segment of the photoreceptors. Photopigments are receptor proteins (opsins) with a prebound chemical agonist (retinal, a derivative of vitamin A). The absorption of light causes a change in the conformation of retinal so that it activates the opsin. This stimulates a G-protein in the disk membrane (transducin) which in turn activates an effector enzyme phosphodiesterase (PDE). PDE breaks down cyclic guanosine monophosphate (cGMP), an intracellular second messenger which keeps the sodium channels in the membrane of the photoreceptor open. The reduction in cGMP causes the sodium channels to close and the photoreceptor membrane to hyperpolarize. (Cf. Kandel, et al., 2000, pp. 507-515; Bear, et al., 2001, pp. 283-299) This is reflected at level L_{P-3} on Figure 2.

[17] Note that here I skip (at least) one intermediate level between the level of the retina and the level of the processes within photoreceptors. At this intermediate level one is able to describe the specific organisation of ganglion cells, amacrine cells, bipolar cells, horizontal cell and photoreceptors forming the retina itself.

Sure, there are further mechanisms explaining e.g. how the electromagnetic radiation is absorbed by the photopigment, or how PDE breaks cGMP down. However, we need not go into further details here. The rough sketch of what happens at these four levels suffices for my present purposes. In what follows, I am going to refer back to this example of how inter-level mechanistic explanation works in order to anchor the abstract points I am about to make.

The notion of constitution plays a central role within the mechanistic framework—as we have seen, the fundamental claim of the approach is that the organised activity of the parts constitutes the whole. However, understanding what is meant by constitution within the mechanistic framework properly is not straightforward.

Note that, on the one hand, proponents of the mechanistic approach evoke the constitution relation as an alternative to identity (cf. e.g. Craver, 2007). To this extent, they are buying in a widely held position. As it is often argued, the lump of clay constitutes the statue, but not *vice versa*: the statue does not constitute the lump of clay. That is, constitution—contrary to identity—is an asymmetric relation (Johnston, 1992; Baker, 1997). On the other hand, though, constitution is usually understood as an intra-level relation (cf. e.g. Paul, 2007). Compare this with the way the mechanistic approach deploys constitution explicitly as an inter-level relation. The lump of clay and the statue are at the same level of composition/aggregation, whereas a whole and its spatially and temporally organised parts are at different levels of composition/aggregation.

Moreover, consider how Craver and Bechtel claim, on the one hand, that it is constitution that relates a higher level with a lower level, whereas, on the other hand, they claim that the relation between levels is symmetrical. They say:

> “The relation is symmetrical precisely because the mechanism as a whole is fully constituted by the organized activities of its parts: a change in the parts is manifest as a change in the mechanism as a whole, and a change in the

> mechanism is also a change in at least some of its component parts." (Craver & Bechtel, 2007, p. 554)

Note, however, that the symmetrical relation in question holds between what the mechanism as a whole does and the behaviour of the spatially and temporally organised parts. That is, the symmetrical relation connects the activity of the whole and the organised activity of the parts. To put it in another way, the proper answer to the question why a certain spatial and temporal organisation of parts constitutes a whole is that because the parts together do the same thing that the whole itself does. That is, the overall behaviour of the parts is *identical* with the behaviour of the whole. Let's unpack this in detail.

Consider what the relation between the behaviour of the whole and the behaviour of the organised structure of its parts is. This question focuses on the causal role played (the activity performed) by the mechanisms as a whole and the causal role played by the spatial and temporal organisation of the parts.

Remember how mechanistic explanations proceed. First, a certain phenomenon is grasped via the task performed by a system, then the system gets decomposed via the identification of its parts, their activity and organisation. The very point of the mechanistic approach is to explain how a system performs certain tasks by understanding how its parts organised in the right way perform the *very same* task. Had the organised structure of the lower level entities performed an activity different from what the higher level entity performs, the account of what happens at the lower level would not have been able to explain the higher level phenomenon in question, and the organised activity of the lower level entities would not have constituted the higher level whole. In this sense, mechanistic explanations *identify* the activity of the spatial and temporal organisation of the parts with the activity of the higher level whole.

That is, the very way mechanistic explanations proceed requires the activity of the organisation of the parts at the lower level to be the same as the activity of the whole at the higher level.

Consider Figure 3, which is a slightly altered version of Figure 1. Figure 3 indicates certain activities of the entities playing part in a mechanistic explanation. Let's say that the entity at the higher level is connected to its context (other entities at its level) by causal relations C_i, C_j, C_k[18], whereas the lower level entities in question are connected to *their* context by causal relations C_l, C_m, C_n. These causal relations characterise the causal role played by the higher level entity and the organised structure of the lower level entities respectively.

The very fact that it is possible to explain the task performed by the higher level entity in terms of the organised activity of the lower level entities entails that the causal roles played at the higher level (C_i, C_j, C_k) and at the lower level (C_l, C_m, C_n) are the same. That is, for a mechanistic explanation to get off the ground the causal profile characterised by C_i, C_j, C_k must be identical with the causal profile characterised by C_l, C_m, C_n. It is thus an internal consequence of the very way mechanistic explanations work that causal connections utilised to characterise the activities at different levels must be identical with each other.

[18] One who is subscribed to Shoemaker's analysis (2003, 2007) might want to say that C_i and C_j represent backward looking whereas C_k represents forward looking causal powers of X_3.

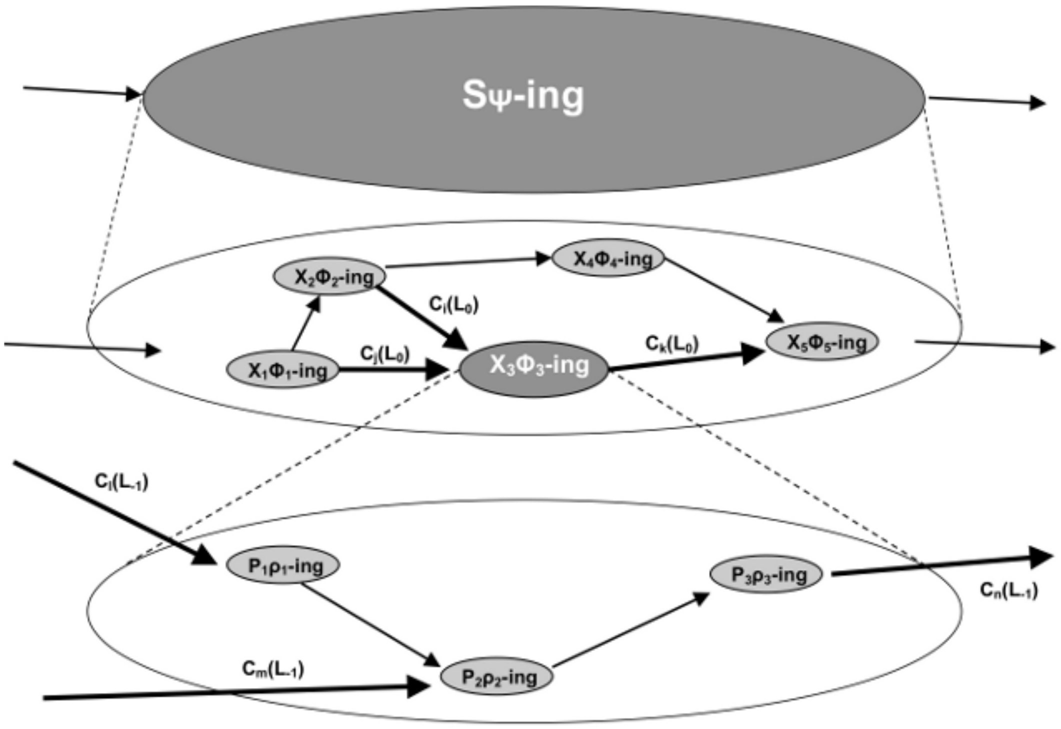

Fig. 2 An inherent commitment of mechanistic explanations. Causal connections C_i, C_j, C_k of entity X_3 belong to the higher level L_0, whereas causal connections C_l, C_m, C_n of the organised structure of entities P_1, P_2, P_3 belong to the lower level L_{-1}. The mechanistic approach must be committed to the claim that the higher level causal role characterised by C_i, C_j, C_k is identical with the lower level causal role characterised by C_l, C_m, C_n.

Note that the organisation of the lower level parts would not constitute the higher level whole if their activities weren't identical. That is, the constitution claim the mechanistic approach makes amounts to an identity claim: the causal role played by the organisation of the parts is the very same causal role that is played by the whole. The conclusion that follows, then, is this: by claiming that the whole is constituted by the organised activity of its parts the mechanistic approach inherently commits itself to the claim that whatever the whole does is something that is done by the organisation of its parts.

There is, however, something that seems to be in tension with the conclusion of the last section: literally, entities at higher levels do different things than entities at lower levels. Consider again the example of the descending levels of mechanisms responsible for the task performed by the eye. E.g. at the level of the lens and the retina entities ‘focus light rays’ and ‘send neuronal signals’, whereas at the lower level of opsin and transducin there are activities like ‘closing ion channels’ and ‘hyperpolarising’. What the claim that at different levels entities do different things emphasises is the simple fact that e.g. whereas at the higher level there are no entities hyperpolarising, at the lower level there is nothing focusing light rays.

Notice how different vocabularies are utilised in order to describe entities and activities at different levels. As one descends from level to level one needs to change the vocabulary of psychology to the vocabulary of anatomy, then to the vocabulary of neuroscience then to that of molecular biology, and so on. In other words, it is hard to see how a task performed at a higher level could possibly be the same as a task performed at a lower level *because* different vocabularies are used at different levels to describe entities and activities.

However, it is possible to reconcile the observation that entities at different levels do different things with the constraint that the mechanistic approach inherently commits itself to the claim that the behaviour of the whole is identical with the overall behaviour of the organisation of its parts. What the mechanistic approach needs here are bridging principles connecting the vocabularies utilised at the different levels. The need for bridging principles is something Bechtel himself is considering as well. He, however, says:

> “Herein lies the explanation for the need for *bridge principles* in the theory-reduction account – different vocabulary is needed to describe what the parts of a mechanism do than is required to describe what the mechanism as a whole does. The appropriate bridge in this case, however, *is not a set of translation rules, but an account of how the operations of the parts of the mechanism are*

> *organized* so as to yield the behavior of the whole mechanism" (Bechtel & Hamilton, 2007, p. 25, emphasis added)

That is, though Bechtel acknowledges that different vocabularies describe activities at different levels, and he also acknowledges that different vocabularies are dealt with by employing bridge principles in the theory-reduction (D-N) account, he still thinks that bridge principles are unnecessary within the mechanistic framework—all one needs is an account of the organisation of the parts' activities.

However, this is not quite right. It is not enough simply to describe the organisation of the constituents. An account of the organisation of the parts' activities is still formulated within the vocabulary of the lower level (e.g. it talks about the spatial and temporal organisation of the transducin activating PDE, the photoreceptor membrane hyperpolarizing, etc.), whereas the behaviours at the higher level are described in another vocabulary used at that particular level (describing how the lens focuses light, the retina converts light into neuronal signal etc.). Proper bridging principles need to be evoked here to connect the different vocabularies. These bridging principles express co-reference of different causal terms: they state e.g. that the term 'the eye transducing light causes the optic nerve to signal' refers to the very same causal role than the term 'the photoreceptor hyperpolarises'. They express that the causal roles evoked by descriptions at different levels are the same.

The moral, then, is this. Mechanistic explanations need to incorporate proper identity statements connecting causal roles (activities) at different levels.[19]

[19] The fact that mechanistic explanations do use identity statements can nicely be pinpointed in some of the texts published by the proponents of the account. For the sake of the example consider the following two quotes.

"In this sketch of events involved in remembering a lecture, I twice stepped down levels by appealing to an *identity* between the effect on a system and a change in constituents of the system. At the lower level the causal story was an ordinary causal one. Then I stepped up a level by appealing to an *identity* between the new operations within the mechanism and the way it behaved as a whole. At the level of the whole the story was again an ordinary causal one." (Bechtel, 2008, pp. 154-155, emphases added)

"And insofar as that non-functioning constitutes the general's death, we explain her death. Notice that when we reach the state of the mechanism that constitutes the state of death, we do not say, with Betty Crocker, that it causes death. *It just is death*." (Craver & Bechtel, 2007, p. 557, emphasis added)

Nowadays, mechanistic explanations are generally considered as our best models of how scientific enquiries connecting micro-phenomena to macro-phenomena are carried out. As we have seen, these enquires typically span through different levels, and rely on different theories at different levels. That is, contrary to what Chalmers and Jackson imply, the 'from-micro-to-macro transition' crosses several levels some of which utilise very different theories with distinct vocabularies—i.e. there is no single language in which both micro and macro descriptions could be formulated.[20] Therefore, in order to be able to cross these levels, one needs some kind of a tool bridging this gap and connecting the terms (descriptions) of the different levels. As we have seen, the transparent version of reductive explanation cannot deliver this tool: there is no *a priori* passage from one level to another. What one needs here, thus, are *a posteriori* identities—identity statements that are prerequisites of the argument generating reductive explanation.

I call these identities 'prior identities', since we need to have them prior to the actual start of the explanatory process. Reductive explanations cannot get off the ground without them.

Prior identities are in many respect quite similar to the identities Hill, McLaughlin, Block and Stalnaker rely on in their non-transparent versions of reductive explanation. However, there is an important dissimilarity: I do not agree with what they claim to be the role of these identities in the process of reductive explanation. My main goal in what follows is to propose a positive account—which is different from theirs—of what the main function of these prior identities is in reductive explanations.

[20] Chalmers and Jackson talk about the language of *physics*. Note that although the particular example of this section is concerned with the micro and macro descriptions of a biological system, the major line of thought generalises: micro descriptions of purely physical systems rely on theories like QCD (quantum chromodynamics), etc. (i.e. the so-called 'standard model' of particle physics), whereas a macro-physical description typically relies on theories like e.g. condensed matter physics with a vocabulary quite *unlike* that of the standard model. (Cf. e.g. Laughlin and Pines (2000) for an illustration of how physicists themselves demarcate macro-physics from micro-physics.)

Before turning our attention to this issue, however, I would like to tie all loose ends by considering Hooker's model of reduction. It is an inherent feature of prior identities that they are not results, but preconditions of successful reductive explanations. I have already argued that the different transparent versions of reductive explanation as they typically appear in the literature related to philosophy of mind are unable to deliver these identities as results of the explanatory process. However, as we have seen it in §6.2 there are other approaches—typically motivated by considerations coming from philosophy of science—whose view on the role identities play in reductions coincides with that of the transparent versions of reductive explanations. Hooker's model of reduction is the most significant example of these approaches (cf. §6.2.2). It argues for a model of reduction which does not require identities (bridge laws) as premises, but rather concludes on them. Therefore, in order to clear the way for my account of prior identities I first need to show that Hooker's model of reduction is unsuccessful in avoiding identities as preconditions. This is what I am going to do in the next part.

7.2.2 Hooker's model of reduction and prior identities

In §6.2.2, I have started my introduction of Hooker's model of reduction with a quote from Ausonio Marras. In that quote Marras argues (in the third point) that the postulation of bridging principles (identities) is just a final step within a reductive attempt which is necessary only for enabling the formal derivation of a law of the 'image-theory' T(R)* from the corresponding law of the original target theory T(R), and thus identities do not play a central role in the process of reduction. What does play the central role is the deduction of the so-called *image* of the target theory from the base theory. And this deduction—since the terms of the image T(R)* are inside the vocabulary of the base T(B)—is straightforward.

In Marras' example we have two equations, namely $pV = 2E/3$ (*L**) and $pV = kT$ (*L*). From this Marras postulates the bridging principle $2E/3 = kT$. First, note that Marras provides only a coarse-grained overview of the process of reduction. It is true that *L* is a law of classical equilibrium thermodynamics, but what one can deduce from the base theory (i.e. statistical mechanics) is not *L**, rather something like the following:

$$(1) \qquad \frac{Nmv^2}{3lwh} \cdot lwh = \frac{2N}{3} \cdot \frac{1}{2} mv^2 .$$

Here *N* is the number, *m* is the mass, *v* is the velocity of the molecules of a gas, and *l*, *h*, *w* denote the length, height and width of a rectangular volume which contains the gas.

Given all this, the following question arises: how can one know that (1) is the so-called *image* of *L*? Marras' coarse-grained story takes it for granted that (1) talks about the same thing as *L*, i.e. that the terms 'pressure' and 'volume' in *L* can be mapped onto an appropriate combination of the terms 'mass', 'velocity', etc. in (1). But remember, *L* is within the vocabulary of T(R) (i.e. classical equilibrium thermodynamics), whereas (1) is within the vocabulary of T(B) (i.e. statistical mechanics), and these two vocabularies are different. This very difference is the reason why one needs bridging principles at all. Hooker's account tries to avoid the need for bridging principles by deducing the image-theory inside the base theory's vocabulary. But to be able to conclude that what is being deduced inside T(B) is an *image* of T(R) one needs connections between the terms of the deduced T(R)* and the original T(R). In Marras' case these connections go as follows:

$$(2) \qquad V \Leftrightarrow l \cdot w \cdot h .$$

$$(3) \qquad p \Leftrightarrow \frac{Nmv^2}{3lwh} .$$

These connections are inter-theoretic connecting principles. They help us in the mapping of the vocabulary of the target theory to the vocabulary of the base theory. Strictly speaking, (2) and (3) express co-reference. What (2) tells us is that the theoretical term '*V*' (volume) of the target theory (classical equilibrium thermodynamics) co-refers with the expression '$l \cdot w \cdot h$' which is constructed out of the terms of the base theory (statistical mechanics). And similarly, (3) tells us that the theoretical term '*p*' (pressure) of thermodynamics co-refers with a certain expression constructed out of some of the theoretical terms of statistical mechanics (i.e. 'velocity', 'mass', etc.). (3) shows that the case in question is indeed a heterogeneous case: one in which there are target-level terms which are not part of the base level vocabulary, and vice versa. Since crossing the levels in heterogeneous cases is the very purpose of introducing bridging principles, these connections are indeed bridging principles—i.e. *prior identities*, which are prerequisites of the process of reduction (in this particular case, the formulation of the 'image-theory').

Of course, Marras is right that in this particular case the postulation of the further bridging principle *2E/3 = kT* is just a final step—the conclusion drawn on the basis of the analogue relation between the target theory and the image-theory,—but he is silent about those bridging principles which connect the terms like 'pressure' of thermodynamics to the so-called analogue terms of statistical mechanics. And these bridging principles are necessary for the construction of the image of T(R), and thus they are preconditions rather than results of the process of reduction.

It is correct to say that Hooker's account is about 'image-deduction', but for the very notion of an 'image' one needs the ability of 'image-recognition'—that is, one needs to know whether some structure within the reducing theory T(B) is an image of the reduced theory T(R). And this ability is ensured by prior identities connecting certain structures of the reducing base theory to corresponding structures of the reduced target theory.

John Bickle, who has advocated a reformulated version of Hooker's original model (Bickle, 1998) might want to argue here that for me (2) and (3) appear as preconditions rather than results only because I forget about the first half of the whole process of reduction. (1), Bickle could argue, can be deduced entirely within the reducing base theory T(B), without the need of any prior identities. Once we have (1), we can *conclude* on (2) and (3) by comparing (1) to (L) and relying on an analogue relation between them. So the whole process of reduction would go like this. First, we deduce (1) from the base theory (i.e. statistical mechanics). This is an image of the target (L). Then by comparing (1) and (L) we can formulate (2) and (3). Once we have them we can further conclude on the famous *2E/3 = kT* identity. That is, all the three identities are in fact consequences of the process of reduction.

This argument, however, is mistaken. To see why, let's reconstruct the deduction of (1) from statistical mechanics (cf. Bickle, 1998, pp. 36-37). Suppose that in a rectangular container (with length *l*, height *h* and width *w*) we have *N* molecules of a gas, each molecule with a mass *m*, and a velocity *v*. Because of the random motion of the molecules, it is a sound assumption that one-third of the molecules are moving along the x-axis. Statistical mechanics tells us that the momentum of a molecule (*I*) is the product of its mass and velocity and that (by assumption) molecular collisions are perfectly elastic. On the basis of this, the total impulse imparted on a wall of the container is:

$$(4) \qquad \sum \Delta I = \frac{N}{3} \cdot 2mv\,.$$

From Newton's Second Law of Motion (5) we can calculate the force exerted by these molecules as the quotient of the total impulse and the time of the total change of momentum. Since the amount of time between two collisions of a molecule with the same wall is *2l/v*:

$$(5) \qquad F = \frac{dI}{dt}\,.$$

$$(6) \qquad F = \frac{N}{3} \cdot 2mv \cdot \frac{v}{2l} = \frac{Nmv^2}{3l}\,.$$

From this Bickle constructs an analogue structure of the Boyle-Charles' law by multiplying the pressure (the quotient of the force and the area of the wall (7)) with the volume of the container:

$$(7) \qquad p = \frac{F}{A}.$$

$$(8) \qquad p \cdot V = \frac{Nmv^2}{3l} \cdot \frac{1}{wh} \cdot lwh = \frac{2N}{3} \cdot \frac{1}{2} mv^2.$$

And finally, comparing (8) with the Boyle-Charles' law Bickle introduces the bridging principles (2) and (3), and consequently *2E/3 = kT*.

On the face of it, this reconstruction supports Bickle's claim. However, in order to construct the analogue structure (8) Bickle relies on (5) and (7). (5) is the definition of force in classical Newtonian mechanics, whereas (7) gives the definition of force in classical thermodynamics. In (8) Bickle uses the term 'force' of Newtonian mechanics as though it was the term 'force' of thermodynamics. But it is not. Bickel is only able to proceed with his deductive argument because he accepts the implicit assumption that the term 'force' of the target theory (thermodynamics) and the term 'force' of the base theory (mechanics) refer to the very same thing. That is, what happens here is an application of yet another bridging principle (i.e. prior identity) connecting a term of Newtonian mechanics to a term of thermodynamics. It is a premise (and not a conclusion) of the argument. Without it the argument would equivocate on the term 'force' and would not go through.

True, some identities are the results of the process of reduction, but there are other, so-called 'prior identities' in play as premises of the same reductive arguments. In other words, one does not only conclude on identities on the basis of comparing image-theories with original theories, but one also needs to rely on certain identities in order to be able to construct the very image-theories themselves.

When one deduces an image-structure in the base theory analogous to a structure of the original target theory one needs some connecting principles to designate what an image-structure would look like within the vocabulary of the base theory. Without these 'prior identities' one would not be able to recognise whether the image deduced was an image of the target structure at all. In the above case, what Bickle does in (8) is replacing the terms of thermodynamics with the terms of Newtonian mechanics within the structure of the Boyle-Charles' law. And what let him do so are the connecting principles between the terms of force and volume of Newtonian mechanics and those of thermodynamics.

One might want to question whether identities are really needed in order to assure us that the term 'force' of thermodynamics and the term 'force' of mechanics refer to the very same thing. Note, however, that the fact that we use the same word in both contexts is quite misleading. The concept 'force' as in mechanics is defined via a change in 'impetus', whereas the concept 'force' as in thermodynamics is defined via the term 'pressure'. Neither 'impetus' is a part of the vocabulary of thermodynamics, nor 'pressure' is part of the vocabulary of mechanics. The very fact that we use the same word reveals the implicit application of an identity claim, which was originally formulated as an empirical hypothesis.

7.2.3 The role of prior identities in reductive explanations

So far I have argued that the transparent model of reductive explanation does not work. Since there is no *a priori* passage available between the target and the base level, one needs to rely on explicit identities as premises of the deductive argument constituting reductive explanation. I call these identities 'prior identities'. It's worth emphasising that despite the somewhat similar name, prior identities are *not a priori* identities—on the contrary, they are *a posteriori*, empirically based identities: instead of being the results of *a priori* conceptual analysis, they are empirical hypotheses, which are necessary premises of any reductive attempt.

Interpreting the identities that play a role in reductive explanations as empirical hypotheses must sound familiar: this is the view defended by Hill (1991), Hill and McLaughlin (1999), Block and Stalnaker (1999), and McLaughlin (2010). According to them, the identities in question provide the best explanation for the observed correlations between the target and the base levels, and thus can be justified on the basis of the method of inference to the best explanation. However, as I shall argue in what follows, this account of what role identities play in reductive explanations is also problematic.

Recall McLaughlin's example quoted in §7.1. He claims that Maxwell made the "bold conjecture" that light waves = electromagnetic waves in order to explain the correlations between the spatial position, the speed in vacuum, and the refractive indices in materials of light waves and electromagnetic waves (cf. McLaughlin, 2010, p. 282). This, however, is an incorrect account of the discovery that light rays are electromagnetic waves. The "bold conjecture" that light rays are electromagnetic waves is not Maxwell's own hypothesis. It was originally proposed by Michael Faraday (1846). Faraday observed that light rays were affected by electromagnetism: under specific conditions a magnetic field could alter the plane of polarisation of a linearly polarised light ray.[21] On the basis of this observation, he subsequently hypothesised that light was a form of electromagnetic vibration. Maxwell knew this hypothesis, and was highly motivated by Faraday's work. In fact, he explicitly acknowledged that Faraday's hypothesis had guided his own research. Right after his famous conclusion drawn from his calculations regarding the velocity of electromagnetic waves, namely that "[t]his velocity is so nearly that of light, that it seems we have strong reason to conclude that light itself [...] is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field

[21] As Maxwell himself summarises Faraday's observation: "When a transparent diamagnetic substance has a ray of polarized light passed through it, and if lines of magnetic force are then produced in the substance by the action of a magnet or of an electric current, the plane of polarization of the transmitted light is found to be changed, and to be turned through an angle depending on the intensity of the magnetizing force within the substance." (Maxwell, 1861, pp. 86-87)

according to electromagnetic laws” (Maxwell, 1865, p. 466) he immediately added that

> “[t]he conception of the propagation of transverse magnetic disturbances to the exclusion of normal ones is distinctly set forth by Professor Faraday in his ‘Thoughts on Ray Vibrations.’ The electromagnetic theory of light, as proposed by him, is the same in substance as that which I have begun to develop in this paper, except that in 1846 there were no data to calculate the velocity of propagation.” (Maxwell, 1865, p. 466)

The moral is that the light = electromagnetic wave identity claim is not evoked in order to explain the correlations McLaughlin cites (cf. same speed, same refractive indices). These correlations had not yet been recognised when the identity claim got originally formulated. What happened, instead, was this: Faraday observed certain similarities between the behaviour of light rays and electromagnetic disturbances (namely that they are similarly affected by magnetic fields). On the basis of this similarity, he proposed an identity claim. This identity claim then motivated Maxwell’s research, who consecutively showed that there is another similarity, namely that the speed of these electromagnetic disturbances in question is “so nearly that of light”. Further research shed light on even more similarities, namely those between the refractive indices of light and electromagnetic waves. Not some kind of inference to the best explanation, but these further evidences uncovering more and more similarities were what finally confirmed the original hypothesis.

As we have seen, proponents of the transparent version of reductive explanation could object here that since, light and electromagnetic waves share all these properties, given that one knows everything there is to know about light rays and about electromagnetic waves, one becomes able to simply infer—and hence does not need to hypothesise—the identity claim connecting them. Against this line of reasoning I have argued that in the general (and philosophically interesting) heterogeneous cases of reductive explanations one cannot realise that entities at the target level and at the base level share certain properties, because target- and base-level entities are described by different vocabularies.[22]

Notice how this claim further complicates the situation for the inference to the best explanation approach. If the target and the base levels are described by different vocabularies, then it is not so straightforward to recognise correlations between the two levels. It is well illustrated by the example of §7.2.2: without prior identities anchoring particular terms from the vocabulary of statistical mechanics to certain terms from the vocabulary of thermodynamics, it is impossible to recognise which structures of statistical mechanics are the 'images' of the structures of thermodynamics, i.e. it is impossible to recognise the very correlations the inference to the best explanation approach starts with.

Perhaps this claim—i.e. that it is exactly the utilisation of prior identities that reveals correlations—is even more evident in the case of mechanistic explanations (cf. §7.2.1). Note that in the local (lower level) context of those entities the organised activity of which constitutes a higher level whole there are other entities which interacts with the supervenience base of the higher level whole. Without prior identities, i.e. without a firm starting point designating certain lower level processes to a higher level process, one has no grounds whatsoever on the basis of which one could start grouping certain lower level entities and activities together as organisations corresponding to higher level activities. That is, without prior identities one could not see the organised activity of which lower level entities it is that co-occurs (i.e. correlates) with a higher level phenomenon. In the case of mechanistic explanations, these prior identities are typically provided by considerations stemming from cross-level experiments—i.e. intervening at the higher level, and recording the effects at the lower level (cf. Woodward, 2003).

[22] Note that in this respect, the above example of light and electromagnetic wave is misleading—from the perspective of the terms 'velocity', 'refraction', etc. it is a special homogeneous case.

Hypothesising prior identities, thus, precedes the proper formulation of correlational claims. That is, explaining correlations is not the primary role prior identities play in reductive explanation. Instead, they are tools to anchor the target level to the base level. They are formulated on the basis of some initial similarities, and then they guide the mapping of target-level phenomena to base-level phenomena. They are justified if they are successful in this process of guiding these mappings, i.e. if on the bases of them, one is able to uncover more and more connections, structural and functional similarities between the target and the base domains. Along these further structural and functional similarities, prior identities make it possible to project the explanatory power of the base level onto the target level, and subsequently, to reductively explain the properties of target level phenomena on the basis of the properties of base level phenomena.

In this latter sense, the model of non-transparent reductive explanation I propose here —which might be called *reductive explanation via prior identities*—resembles Block and Stalnaker's version of non-transparent reductive explanation. As we have seen in §6.3.3, Block and Stalnaker emphasise that identities make it possible to transfer explanatory and causal power from the base level to the target level. Though in this respect, my approach is similar to theirs, there are also significant differences between the two view. Most importantly, Block and Stalnaker rely on the principle of inference to the best explanation in justifying the identities themselves. They claim that we accept identities over mere correlations, exactly because identities allow us to transfer explanations from the base level to the target level. That is, Block and Stalnaker's approach concentrates on the advantages of identity claims over correlational claims. This attitude suggests that identities are evoked only when correlations are observed. Contrary to this, my approach draws attention to the fact that certain prior identities are necessary even for formulating correlational claims. Moreover, these prior identities do not compete with correlational claims, but with other, alternative prior identities—alternative ways of mapping the target level onto the base level. Those prior identities are supported over the alternatives which allow for an extension of the initial mapping, and make the process of uncovering further similarities between the target and the base domains possible.

The closest match to my account of reductive explanation via prior identities is the so-called *heuristic identity theory* of McCauley and Bechtel (McCauley, 1981; Bechtel & McCauley, 1999; McCauley & Bechtel, 2001). McCauley and Bechtel are mainly concerned with the role identities play in scientific explanations in general. They observe that different scientific theories describing the target and base levels of typical reductive explanations usually co-evolve: the explanatory tools and devices developed at one level affect and foster novel methods, models and hypotheses at the other level, and vice versa. This co-evolution is due to the formulation of heuristic identities, which are hypothesised in order to connect the different levels and thus advance explanatory endeavours and scientific development. For example, psychologists have developed certain behavioural tools, which then guided the discovery of "hypotheses about underlying neural mechanisms", which, then, in turn prompted "the development of increasingly sophisticated information-processing models at the psychological level" (McCauley & Bechtel, 2001, p. 745). The main driving force of this mutual influence "is the hypothetical identifications of information-processing activities with brain processes (in characteristic brain areas)." (McCauley & Bechtel, 2001, p. 745)

McCauley and Bechtel think of heuristic identities as crucial devices of furthering the advancement of science. As they put it:

> "Hypothesizing cross-scientific identities is a pivotal engine of scientific development. Hypothetical identities in interlevel contexts serve as valuable heuristics of discovery for inquiry at both of the explanatory levels involved. Crucially, scientists accept or reject these hypotheses for the same reasons that they accept or reject any other hypothesis in science. These are the same reasons involved in establishing the truth of any abductive inference, namely the resulting hypotheses' abilities to stand up to empirical evidence, to stimulate new research and to foster the integration of existing knowledge." (McCauley & Bechtel, 2001, p. 751)

That is, McCauley and Bechtel's heuristic identities, very similar to my prior identities, are justified on the basis of their success in mapping the higher and the lower levels onto each other, thus allowing for the projection of explanatory power originally present at one level onto the other level.

McCauley and Bechtel go on and identify the main role of hypothesising these heuristic identities as follows:

> "According to [the heuristic identity theory], [...] identities are not the conclusions of scientific research but the premises. The logic behind their use looks to the converse of Leibniz's law. Instead of appealing to the identity of indiscernibles, this strategy exploits the indiscernibility of identicals. [...] The theories at each level ascribe distinct properties to the entities and processes that the interlevel, hypothetical identities connect. Since they both address features of the same physical systems, though, scientists have grounds from the outset to expect that these accounts will gradually evolve so as to mirror one another more and more. By virtue of the proposed identities, scientists can use related research at each explanatory level to stimulate discovery at the other." (McCauley & Bechtel, 2001, pp. 753-754)

Again, heuristic identities within McCauley and Bechtel's approach of scientific research play a role very similar to the role that prior identities play in my approach of reductive explanation. They both are premises, starting points, which are necessary for specific kinds of scientific research or reductive explanation to get off the ground. Heuristic identities make it possible to use the results of certain research originally related to one level at another level, and similarly, prior identities make it possible to anchor the target level phenomenon to the base domain and utilise the explanatory apparatus of the base domain to account for the target phenomenon. Moreover, just as heuristic identities are evaluated on the basis of how efficient they are in fostering the co-evolution of the two different accounts originally formulated within the two distinct scientific theories to mirror each other, prior identities are evaluated on the basis of how efficient they are in mapping the target and the base domains onto each other and advancing the discovery of further similarities.

Finally, McCauley and Bechtel's heuristic identity theory is similar to my account of reductive explanation via prior identities in one more respect: heuristic identities, just like prior identities are *unexplained explainers*—they themselves cannot be explained in the sense proponents of transparent reductive explanation think identities are explained, but rather are hypothesised on the basis of empirical considerations, and utilised in explaining certain features of phenomena at one level in terms of the features of phenomena at the other level (in the case of reductive explanation, the target level and the base level respectively). As McCauley and Bechtel puts is:

> "What matters about hypothetical cross-scientific identities is not how they should be explained (they can't be) but what they explain, how they suggest (and contribute to) other, empirically successful, explanatory hypotheses, and how they create opportunities for scientists at one explanatory level to enlist methods and evidence from alternative levels of explanation." (McCauley & Bechtel, 2001, pp. 756-757)

7.3 Conclusion: the Monadic Marker Account as a Reductive Explanation Via Prior Identities

In the last part of this dissertation (Chapter 6 and Chapter 7) I have investigated the process of reductive explanation. Two general types of reductive explanation have been distinguished, the so-called transparent and the so-called non-transparent versions. I have argued that the transparent version of reductive explanation (a.k.a. the *a priori* passage view) faces a dilemma, neither horns of which is acceptable for the proponents of the transparent version. I have also shown that the traditional understandings of the non-transparent version misrepresent the role identities play in reductive explanations.

On the basis of how actual reductive attempts proceed, I have proposed a new account of reductive explanation, which utilises so-called prior identities. *Reductive explanation via prior identities*, thus, is a novel version of the non-transparent model.

It relies on identities as premises in the process of deducing a target-level claim from base level theories and descriptions. These identities are prerequisites rather than outcomes of successful reductive explanations—hence their name, *prior identities*. Prior identities are proper identity claims, which themselves are unexplained but are nevertheless necessary for mapping the features to be explained onto the features the explanation relies on. That is, they designate which base level phenomenon corresponds to which target level phenomenon, thereby projecting the explanatory power of the base level onto the target level. They are hypothesised in order to foster the formulation of explanatory claims accounting for target level phenomena in terms of base level processes—and they are justified if they are successful in doing just that. Prior identities are justified if they help projecting base level explanations to 'new territories' of the target level, covering phenomena, which would have otherwise (i.e. had the prior identities not been formulated) remained unexplained in lower level terms.

Now recall the identity claim proposed in Chapter 4 as part of the Monadic Maker Account of conscious experience. It says that the phenomenal character of simple conscious experiences is identical with how monadic markers are interpreted by central processes of a cognitive system. First, note that formulating this identity claim was not an ad hoc step: it has been hypothesised on the basis of certain similarities recognised between the phenomenal domain and the cognitive-representational domain.[23] Second, note that it has not been used to explain the observed similarity itself, but rather to extend the initial mapping of the phenomenal domain to the target domain and uncover further similarities between these two domains—e.g. similarities between certain features of monadic markers, and features of secondary qualities like their 'dissimilarity' to physical properties, or features of phenomenal qualities like their functional un-analysability. Third, note that on the basis of these further similarities uncovered with the aid of the initial identity claim, explanations of certain characteristics of the phenomenal domain in terms of

[23] Cf. the similarities between the observations (O1), (O2), (O3) and the claims (C1*), (C2*), (C3*) in §4.3.3.

characteristics of the cognitive-representational domain have been proposed. In particular, the identity claim of the Monadic Marker Account made it possible to put forth explanations of the primary-secondary quality distinction (cf. §5.3.1), the fact that conscious experiences resist functionalisation (cf. §5.3.2), and why the explanatory gap arises, why Mary learns something new, or why zombies seem conceivable (cf. §5.3.3)—i.e. to formulate explanations of target-level phenomena solely in terms of the base level.

That is, the identity claim of the Monadic Marker Account is a prior identity. It is an unexplained explainer, deployed in order to uncover similarities between the phenomenal and the cognitive-representational domains, and to project the explanatory power of the cognitive-representational domain onto the phenomenal domain. It helps explaining features and phenomena of the phenomenal domain which otherwise would remain unexplained. These explanations, thus,—in accordance with the model of reductive explanation via prior identities, and contrary to the received view of contemporary literature—are reductive explanations *proper*. The identity claim featuring in them is justified similarly to the way standard scientific identities are justified, and the explanations provided are formulated analogously to how standard scientific explanations are formulated.

The Monadic Marker Account, then, is the result of the very same methodology that is characteristic of all those scientific endeavours, which aim at fostering understanding of the reductive kind—i.e. the understanding of particular target phenomena in terms of certain lower level base phenomena. In this sense, the Monadic Marker Account introduces the problem of phenomenal consciousness into scientific discourse, and therefore offers a bridge between the philosophy and the science of consciousness: it offers an approach to conscious experience which, on the one hand, tries to account for the philosophically most important features of consciousness, whereas, on the other hand, does it in a way which smoothly fits into the everyday practice of scientific research.

The Monadic Marker Account provides a theoretical framework, a cognitive model, which when implemented by a particular empirical theory—originally focusing on those aspects of consciousness that can be operationalised—helps the embedding theory to address important questions related to the qualitative character of conscious experience. Ned Block (2010) famously criticises cognitive and biological theories of consciousness that they are unable to account for the presence of the explanatory gap. The novel approach promoted in my dissertation holds the promise of being helpful in just that—it can act as a universal plug-in for different scientific theories of consciousness and further their reach by providing access to its own explanatory resources connecting cognitive processes to the distinguishing features of conscious experience.[24]

[24] One might feel tempted here to raise a final challenge at this point by drawing attention to the fact that the conclusion of Chapter 5 and the conclusion of Chapter 7 seem to be in tension with each other. For, on the one hand, in Chapter 4 and Chapter 5 it has been acknowledged that there is an explanatory gap between the phenomenal and the physical domains, whereas, on the other hand, Chapter 7 has concluded that features of conscious experience can be explained very similarly to how standard scientific explanations proceed, which seems to imply that, in fact, there is no explanatory gap between the phenomenal and the physical domains. To answer this challenge, note first that the explanatory gap that has been acknowledged in Chapter 4 and Chapter 5 is a gap between a particular phenomenal quality and a corresponding monadic marker. The Monadic Marker Account does not close this gap—it does not provide any explanation of why a particular monadic marker when interpreted by a cognitive system properly embedding it is identical with a specific phenomenal quality. Rather, the Monadic Marker Account, presupposing such identities, provides explanations of, for example, why primary qualities resemble physical properties whereas secondary qualities don't, why consciousness resists functionalisation, and why the explanatory gap between a particular monadic marker and phenomenal quality pair arises. That is, whereas the acknowledged explanatory gap claims about a particular target phenomenon that it is unexplainable in base level terms, the proposed reductive explanation explains a different target phenomenon. Second, note the related point that the novel account proposed here, i.e. reductive explanation via prior identities, takes the identity claims featuring in this approach to be unexplained explainers—they are not conclusions of the deductive arguments yielding reductive explanations, but rather premises in such arguments, and are justified on independent grounds. That is, in the sense in which the transparent version of reductive explanation uses the term, there are indeed explanatory gaps between the entities flanking these identity signs. But again, the reductive explanations proposed do not explain these identities, but rather utilise them to explain other phenomena. However, in accordance with the reductive explanation via prior identities view, this is true of all cases of scientific reductions. So why does the case of explaining the features of consciousness seem intuitively different from the case of, say, explaining the features of water? On the present account, there are similar explanatory gaps between the features of water and features of H_2O than between monadic markers and phenomenal qualities. Why do these two cases nevertheless seem to be different? This intuitive difference, I propose, stems from the functional un-analysability of phenomenal qualities. The functional un-analysability of phenomenal qualities is a unique feature of theirs, which tell them apart from all the properties of physical systems. Although this difference does not affect the reductive explanations provided (cf. Chapter 7), it gives rise to the aforementioned 'intuition of distinctness'. This however, does not pose a problem for the Monadic Marker Account—since it is able to explain functional un-analysability, it is also able to explain the intuition of distinctness. Cf. Papineau (2002, 2010) for arguments that the intuition of distinctness plays a crucial role in the mind-body issue.

Bibliography

Alexander, S. (1920). *Space, Time, and Deity*. London: Macmillan.

Armstrong, D. (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.

Aydede, M., & Güzeldere, G. (2005). Cognitive Architecture, Concepts, and Introspection: An Information-Theoretic Solution to the Problem of Phenomenal Consciousness. *Nous, 39*(2), 197-255.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Baddley, A., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47-89). New York: Academic Press.

Baker, L. R. (1997). Why Constitution is not Identity. *The Journal of Philosophy, 94* (12), 599-621.

Balog, K. (2009). Phenomenal Concepts. In B. P. McLaughlin (Ed.), *The Oxford Handbook of Philosophy of Mind* (pp. 292-312). Oxford: Oxford University Press.

Balog, K. (2012a). Acquaintance and the Mind-Body Problem. In S. Gozzano & C. Hill (Eds.), *New Perspectives on Type Identity: The Mental, the Physical* (pp. 16-42). Cambridge: Cambridge University Press.

Balog, K. (2012b). In Defense of the Phenomenal Concept Strategy. *Philosophy and Phenomenological Research, 84*(1), 1-23.

Bates, J. (2009). A defense of the explanatory argument for physicalism. *Philosophical Quarterly, 59*(235), 315-324.

Batterman, R. (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*: Oxford University Press.

Baumgartner, G. (1960). Indirekte Größenbestimmung der rezeptiven Felder der Retina beim Menschen mittels der Hermannschen Gittertäuschung. *Pflügers Archiv für die gesamte Physiologie, 272*, 21-22.

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2001). *Neuroscience: Exploring the Brain*. Baltimore: Lippincott Williams & Wilkins.

Bechtel, W. (2007). Reducing psychology while maintaining its autonomy via mechanistic explanation. In M. Schouten & H. Looren de Jong (Eds.), *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction* (pp. 172-198). Oxford: Basil Blackwell.

Bechtel, W. (2008). *Mental mechanisms: philosophical perspectives on cognitive neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W., & Hamilton, A. (2007). Reduction, integration, and the unity of the sciences. In T. Kuipers (Ed.), *Philosophy of Science: Focal Issues* (Vol. 1 of the Handbook of the Philosophy of Science). New York: Elsevier.

Bechtel, W., & McCauley, R. N. (1999). Heuristic identity theory (or back to the future): The mind-body problem against the background of research strategies in cognitive neuroscience. In M. Hahn & S. C. Stones (Eds.), *Proceedings of* *the twenty-first meeting of the Cognitive Science Society* (pp. 67-72). Mahwah, NJ: Erlbaum.

Beckermann, A. (2000). The perennial problem of the reductive explainability of phenomenal consciousness - C.D. Broad on the explanatory gap. In T. Metzinger (Ed.), *Neural Correlates of Consciousness - Empirical and conceptual questions* (pp. 41-55). Cambridge, MA: MIT Press.

Bedau, M. (1997). Weak Emergence. *Philosophical Perspectives, 11*, 375-399.

Bedau, M. (2008). Is Weak Emergence Just in the Mind? *Minds and Machines, 18* (4), 443-459.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115-147.

Biederman, I., & Shiffrar, M. (1987). Sexing day-old chicks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 640-645.

Bird, A. (2007). *Nature's Metaphysics*. Oxford: Oxford University Press.

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.), *Readings in the Philosophy of Psychology* (Vol. 1., pp. 268-305). Cambridge: Harvard University Press.

Block, N. (1995). On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences, 18*, 227-248.

Block, N. (2007). Max Black's Objection to Mind-Body Identity. In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 249-306). Oxford: Oxford University Press.

Block, N. (2010). Comparing the Major Theories of Consciousness. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences, 4th edition* (pp. 1-12). Boston: MIT Press.

Block, N. (forthcoming). Functional Reduction. In T. Horgan, D. Sosa & M. Sabates (Eds.), *Supervenience in Mind: A Festschrift for Jaegwon Kim*. Cambridge, MA: MIT Press.

Block, N., & Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review, 108*, 1-46.

Boogerd, F. C., Bruggeman, F. J., Richardson, R. C., Stephan, A., & Westerhoff, H. V. (2005). Emergence and Its Place in Nature: A Case Study of Biochemical Networks. *Synthese, 145*(1), 131-164.

Boynton, R. (1997). Insights gained from naming the OSA colours. In C. Hardin & L. Maffi (Eds.), *Colour categories in thought and language*. Cambridge, MA: MIT Press.

Broad, C. D. (1925). *The Mind and Its Place in Nature*. London: Routledge and Kegan Paul.

Brook, A., & Raymont, P. (2010). The Unity of Consciousness. *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/fall2010/entries/consciousness-unity/>

Bryson, J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 165-190.

Campbell, F., & Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology, 197*, 551-566.

Campbell, K. K. (1970). *Body and Mind*. New York: Doubleday.

Carey, B. (2011). Overdetermination And The Exclusion Problem. *Australasian Journal of Philosophy, 89*(2), 251-262.

Carruthers, P. (2000a). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Carruthers, P. (2000b). The evolution of consciousness. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the Human Mind* (pp. 254-275). Cambridge: Cambridge University Press.

Carruthers, P. (2001). Consciousness: explaining the phenomena. In D. Walsh (Ed.), *Naturalism, Evolution and Mind* (pp. 61-86). Cambridge: Cambridge University Press.

Carruthers, P. (2004). Phenomenal Concepts and Higher-Order Experiences. *Philosophy and Phenomenological Research, 68*(2), 316-336.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Chalmers, D. (1995). Facing up to the problem of Consciousness. *Journal of Consciousness Studies, 2*, 200-219.

Chalmers, D. (1996). *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press.

Chalmers, D. (2002). Does Conceivability Entail Possibility? In T. Szabo Gendler & J. Hawthorne (Eds.), *Conceivability and Possibility* (pp. 145-200). Oxford: Oxford University Press.

Chalmers, D. (2003). Consciousness and its place in nature. In S. Stich & T. Warfield (Eds.), *Blackwell Guide to the Philosophy of Mind* (pp. 102-142). Oxford: Blackwell.

Chalmers, D. (2004). The Foundations of Two-Dimensional Semantics. In M. Garcia-Caprintero & J. Macia (Eds.), *Two-Dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press.

Chalmers, D. (2007). Phenomenal Concepts and the Explanatory Gap. In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 167-194). Oxford: Oxford University Press.

Chalmers, D. (2010a). *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, D. (2010b). The Two-Dimensional Argument against Materialism. In D. Chalmers (Ed.), *The Character of Consciousness* (pp. 141-205). Oxford: Oxford University Press.

Chalmers, D., & Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review, 110*, 315-361.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy, 78*, 67-90.

Churchland, P. M. (1985). Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy, 82*, 8-28.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/ Brain*. Cambridge, MA: MIT Press.

Conman, J. (1968). On the Elimination of Sensations and Sensations. *Review of Metaphysics, 22*, 15-35.

Craik, F., & Lockhart, R. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Crane, T. (1991). All God has to do. *Analysis, 51*, 235-244.

Crane, T. (2001a). *Elements of mind : an introduction to the philosophy of mind*. Oxford: Oxford University Press.

Crane, T. (2001b). The Significance of Emergence. In C. Gillet & B. Loewer (Eds.), *Physicalism and its Discontents* (pp. 207-224). Cambridge: Cambridge University Press.

Crane, T., & Mellor, D. H. (1990). There is no Question of Physicalism. *Mind, 99*, 185-206.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science, 68*(1), 53-74.

Craver, C. F. (2007). *Explaining the brain: mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon Press

Craver, C. F., & Bechtel, W. (2007). Top-down Causation Without Top-down Causes. *Biology & Philosophy, 22*(4), 547-563.

Crick, F., & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature, 375*, 121-123.

Davidson, D. (1970). Mental Events. In L. Foster & J. Swanson (Eds.), *Experience and Theory* (pp. 79-101). Amherst: University of Massachuusetts Press.

de Gardelle, V., & Kouider, S. (2009). Cognitive theories of consciousness *Encyclopedia of Consciousness*: Elsevier.

de Gardelle, V., Sackur, J., & Kouider, S. (2009). Perceptual illusions in brief visual presentations. *Consciousness and Cognition, 18*(3), 569-577.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204-211.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79* (1-2), 1-37.

Dennett, D. C. (1991). *Consciousness explained* (1st ed.). Boston: Little, Brown and Co.

Descartes, D. (1984). *The Philosophical Writings of Descartes* (J. Cottingham, R. Stoothoff & D. Murdoch, Trans.). Cambridge: Cambridge University Press.

Diaz-Leon, E. (2008). Defending the Phenomenal Concept Strategy. *Australasian Journal of Philosophy, 86*(4), 597-610.

Diaz-Leon, E. (2010). Can Phenomenal Concepts Explain The Epistemic Gap? *Mind, 119*(476), 933-951.

Diaz-Leon, E. (2011). Reductive explanation, concepts, and a priori entailment. *Philosophical Studies, 155*, 99-116.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge: Bradford Books.

Elder, C. (1994). Laws, Natures, and Contingent Necessities. *Philosophy and Phenomenological Research, 54*, 649–667.

Ellis, B. (2001). *Scientific Essentialism*. Cambridge: Cambridge University Press.

Faraday, M. (1846). Thoughts on Ray Vibrations. *Philosophical Magazine, 28*, 447-452.

Fazekas, P., & Zemplén, G. (2005). Molyneux's Questions - from the philosophical problem of representation to intermodal transfer (in Hungarian). *Magyar Pszichológiai Szemle, 60*(4), 527-552.

Feigl, H. (1958). The "Mental" and the "Physical". In H. Feigl, M. Scriven & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press.

Fodor, J. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese, 28*, 77-115.

Fodor, J. (1983). *The Modularity of the Mind*. Cambridge, MA: MIT Press.

Fodor, J. (1987). Why there still has to be a language of thought *Psychosemantics* (pp. 135-154). Cambridge, MA: MIT Press.

Fodor, J. (1998). *Concepts*. Cambridge, Mass.: MIT Press.

Fodor, J. (2008). *LOT2 – The Language of Thought Revisited*. Oxford: Clarendon Press.

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis, 23*, 121-123.

Gillett, C., & Witmer, G. (2001). A "Physical Need": Physicalism and the via Negativa. *Analysis, 61*, 302-309.

Hardin, C. (1988). *Color for Philosophers: Unweaving the Rainbow*. Indianapolis, MA: Hackett.

Harman, G. (1966). The Inference to the Best Explanation. *Philosophical Review, 74*, 88-95.

Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives, 4*, 31-52.

Hawthorne, J. (2001). Causal Structuralism. *Philosophical Perspectives, 15*, 361-378.

Hawthorne, J. (2002). Blocking Definitions of Materialism. *Philosophical Studies, 110*(2), 103–113.

Hayworth, K. J., & Biederman, I. (2006). Neural evidence for intermediate representations in object recognition. *Vision Research, 46*(23), 4024-4031.

Held, R., Ostrovsky, Y., deGelder, B., Gandhi, T., Ganesh, S., Mathur, U., et al. (2011). The newly sighted fail to match seen with felt. *Nature Neuroscience*, published online 10 April 2011.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Hempel, C. G. (1969). Reduction: Ontological and Linguistic Facets. In S. Morgenbesser, P. Suppes & M. White (Eds.), *Philosophy, Science, and*

Method: Essays in Honor of Ernest Nagel (pp. 179-199). New York: St. Martin's Press.

Hermann, L. (1870). Eine Erscheinung simultanen Contrastes. *Pflügers Archiv für die gesamte Physiologie, 3*, 13-15.

Hill, C. S. (1991). *Sensations*. Cambridge: Cambridge University Press.

Hill, C. S., & McLaughlin, B. P. (1999). There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research, 59*, 445–454.

Hooker, C. (1981). Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorical Reduction. *Dialogue, 20*, 38-57, 201-236, 496-529.

Horgan, T. (1983). Supervenience and Microphysics. *Pacific Philosophical Quarterly, 63*, 29–43.

Horgan, T. (1984). Jackson on Physical Information and Qualia. *The Philosophical Quarterly, 34*, 147-152.

Horgan, T. (1993). From Supervenience to Superdupervenience: Meeting the Demands of a Material World. *Mind, 102*(408), 555-586.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480-517.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly, 32*, 127-136.

Jackson, F. (1986). What Mary Didn't Know. *The Journal of Philosophy, 83*, 291-295.

Jackson, F. (1994). Armchair Metaphysics. In J. O'Leary Hawthorne & M. Michael (Eds.), *Philosophy in Mind* (pp. 23-42). Dordrecht: Kluwer.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson, F. (2003). From H2O to Water: The Relevance to *A Priori* Passage. In H. Lillehammer & G. Rodriguez-Pereyra (Eds.), *Real Metaphysics*. New York: Routledge.

Jakab, Z. (2000). Ineffability of Qualia: A Straightforward Naturalistic Explanation. *Consciousness and Cognition, 9*, 329-351.

Jakab, Z. (2003). Phenomenal Projection. *Psyche, 9*(4), 1-12.

Jakab, Z. (2006). Revelation and Normativity in Visual Experience. *Canadian Journal of Philosophy, 36*(1), 25-56.

Johnston, M. (1992). Constitution is not Identity. *Mind, 101*, 89-105.

Kallestrup, J. (2011). Supervenience. *Oxford Bibliographies Online: Philosophy*. Oxford: Oxford University Press.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science*. New York: McGraw-Hill.

Kaplan, D. (1989). Demonstratives – An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals. In J. Almog, J. Perry & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481-564).

Kauffman, S., & Clayton, P. (2006). On emergence, agency, and organization. *Biology & Philosophy, 21*(4), 501-521.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu Rev Psychol, 55*, 271-304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Curr Opin Neurobiol, 13*(2), 150-158.

Kim, J. (1993a). *Supervenience and Mind: Selected Philosophical Essays*. London: Cambridge University Press.

Kim, J. (1993b). The Non-Reductivist's Troubles with Mental Causation. In J. Heil & A. Mele (Eds.), *Mental Causation* (pp. 189-210). Oxford: Clarendon Press.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge: Bradford Books.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies, 95*, 3-36.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kim, J. (2006a). Being Realistic about Emergence. In P. Clayton & P. Davies (Eds.), *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion* (pp. 189-202). New York: Oxford University Press.

Kim, J. (2006b). Emergence: Core ideas and issues. *Synthese, 151*(3), 547-559.

Kim, J. (2009). "Supervenient and yet not deducible": Is there a coherent concept of ontological emergence? In A. Hieke & H. Leitgeb (Eds.), *Reduction: Between the Mind and the Brain* (pp. 53-72): Ontos Verlag.

Kirk, R. (1974a). Sentience and Behaviour. *Mind, 83*, 43-60.

Kirk, R. (1974b). Zombies versus Materialists. *Aristotelian Society, 48*, 135-152.

Kirk, R. (1996). Physicalism Lives. *Ratio, 9*, 85-89.

Koch, C., & Braun, J. (1996). Towards a neuronal correlate of visual awareness. *Current Opinion in Neurobiology, 6*, 158-164.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10*, 319-326.

Kosslyn, S. (1994). *Image and brain*. Cambridge: MIT Press.

Kouider, S. (2009). Neurobiological theories of consciousness *Encyclopedia of consciousness*: Elsevier.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences, 14*(7), 301-307.

Kouider, S., & Dupoux, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. *Psychological science : a journal of the American Psychological Society / APS, 15*(2), 75-81.

Kripke, S. (1980). *Naming and Necessity*. Cambridge: Harvard University Press.

Kulvicki, J. (2004). Isomorphism in Information Carrying Systems. *Pacific Philosophical Quarterly, 85*, 380-395.

Kulvicki, J. (2005). Perceptual Content, Information, and the Primary/Secondary Quality Distinction. *Philosophical Studies, 122*, 103-131.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494-501.

Land, E. H. (1964). The retinex. *Scientific American, 52*, 247-264.

Laughlin, R. B., & Pines, D. (2000). The Theory of Everything. *Proceedings of the National Academy of Sciences, 97*(1), 28-31.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis, 20*(7), 1434-1448.

Leibniz, G. (1898). *The Monadology* (R. Latta, Trans.). Oxford: Oxford University Press.

Levin, J. (2007). What is a Phenomenal Concept? In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 87-110). Oxford: Oxford University Press.

Levin, J. (2008). Taking Type B Physicalism seriously. *Mind & Language, 23*, 402-425.

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly, 64*, 354-361.

Levine, J. (1993). On leaving out what it's like. In M. Davies & G. Humphreys (Eds.), *Consciousness: Psychological and Philosophical Essays* (pp. 121-136). Oxford: Blackwell.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.

Levine, J. (2007). Phenomenal Concepts and the Materialist Constraint. In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 145-166). Oxford: Oxford University Press.

Lewes, G. H. (1875). *Problems of Life and Mind*. London: Kegan Paul, Trench, Turbner, and Co.

Lewis, D. (1983a). Extrinsic properties. *Philosophical Studies, 44*, 197-200.

Lewis, D. (1983b). New Work for a Theory of Universals. *Australasian Journal of Philosophy, 61*, 343-377.

Lewis, D. (1990). What experience teaches. In W. Lycan (Ed.), *Mind and Cognition: A Reader*. Oxford: Blackwell.

Lewis, D. (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy, 106*(9), 475-502.

Lloyd, M. C. (1925). Emergent evolution. *Mind, 34*(133), 70-74.

Loar, B. (1990). Phenomenal States. In J. Tomberlin (Ed.), *Philosophical Perspectives* (pp. 81-108.). Northridge: Ridgeview Publishing Company.

Loar, B. (1997). Phenomenal States. In N. Block, O. Flanagan & G. Güzeldere (Eds.), *The Nature of Consciousness* (pp. 597-616). Cambridge: The MIT Press.

Locke, J. (1690/1987). *An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Loewer, B. (2001). From Physics to Physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and Its Discontents* (pp. 37-56). Cambridge: Cambridge University Press.

Lycan, W. (2008). Representational Theories of Consciousness. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/fall2008/entries/consciousness-representational/>

Lycan, W., & Pappas, G. (1972). What is Eliminative Materialism? *Australasian Journal of Philosophy, 50*, 149-159.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Mamassian, P. (2006). Bayesian inference of form and shape. *Progress in Brain Research, 154*, 265-270.

Margolis, E. (1998). How to Acquire a Concept. *Mind and Language, 13*(3), 347-369.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Marras, A. (2002). Kim on Reduction. *Erkenntnis, 57*, 231-257.

Maxwell, J. C. (1861). On Physical Lines of Force (Part IV). *Philosophical Magazine, March 1861*, 85-95.

Maxwell, J. C. (1865). A Dynamical Theory of the Electromagnetic Field. *Philosohpical Transactions of the Royal Society of London, 155*, 459-512.

McCauley, R. N. (1981). Hypothetical identities and ontological economizing: Comments on Causey's program for the unity of science. *Philosophy of Science, 48*(218-227).

McCauley, R. N., & Bechtel, W. (2001). Explanatory Pluralism and Heuristic Identity Theory. *Theory & Psychology, 11*(6), 736-760.

McDermott, D. (2001). *Mind and mechanism*. Cambridge: MIT Press.

McIntyre, L. (1999). The emergence of the philosophy of chemistry. *Foundations of Chemistry, 1*(1), 57-63.

McIntyre, L. (2007). Emergence and reduction in chemistry: ontological or epistemological concepts? *Synthese, 155*(3), 337-343.

McLaughlin, B. P. (1992). The Rise and Fall of British Emergentism. In A. Beckermann, H. Flohr & J. Kim (Eds.), *Emergence or Reduction? Essays on the Prospects of Non-reductive Physicalism* (pp. 49-93). Berlin: De Gruyter.

McLaughlin, B. P. (1997). Emergence and Supervenience. *Intellectica, 2*, 25-43.

McLaughlin, B. P. (2001). In Defense of New Wave Materialism: A Response to Horgan and Tienson. In C. Gillett & B. Loewer (Eds.), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

McLaughlin, B. P. (2010). Consciousness, Type Physicalism, and Inference to the Best Explanation. *Philosophical Issues, 20*(1), 266-304.

Mellor, D. H. (1995). *The Facts of Causation*. London: Routledge.

Melnyk, A. (1997). How to Keep the Physical in Physicalism. *The Journal of Philosophy, 94*, 622-637.

Menzies, P., & List, C. (2010). The Causal Autonomy of the Special Sciences. In C. MacDonald & G. MacDonald (Eds.), *Emergence in Mind* (pp. 108-128). Oxford: Oxford University Press.

Metzger, W. (1930). Optische Untersuchungen am Ganzfeld. *Psychologische Forschung, 13*, 6-29.

Metzinger, T. (1995). Faster than Thought: Holism, Homogeneity and Temporal Coding. In T. Metzinger (Ed.), *Conscious Experience* (pp. 425-464). Thorverton: Imprint Academic.

Metzinger, T., & Walde, B. (2000). Commentary on Jakab's "Ineffability of Qualia". *Consciousness and Cognition, 9*(3), 352-362.

Mill, J. S. (1843). *A System of Logic*. London: Longmans, Green, Reader, and Dyer.

Montero, B. (2001). Post-Physicalism. *Journal of Consciousness Studies, 8*, 61-80.

Montero, B., & Papineau, D. (2005). A defence of the *via negativa* argument for physicalism. *Analysis, 65*, 233-237.

Morgan, L. (1923). *Emergent Evolution*. London: Williams & Norgate.

Nagel, E. (1961). *The Structure of Science*. London: Routledge & Kegan Paul.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 4*, 81-108.

Nagel, T. (1979). *Mortal Questions*. Cambridge: Cambridge University Press.

Nemirow, L. (1980). Review of Thomas Nagel, Mortal Questions. *Philosophical Review, 89*, 473-477.

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In W. Lycan (Ed.), *Mind and Cognition: A Reader* (pp. 490–499). Oxford: Blackwell.

Neurath, O. (1931a). Physicalism: The Philosophy of the Vienna Circle. *The Monist, 41*, 618-623.

Neurath, O. (1931b). Physikalismus. *Scientia*(November), 297-303.

Ney, A. (2008). Physicalism as Attitude. *Philosophical Studies, 138*, 1-15.

Nida-Rümelin, M. (1996). What Mary couldn't know. In T. Metzinger (Ed.), *Phenomenal Consciousness*. Schoenigh: Paderborn.

Nida-Rümelin, M. (2010). Qualia: The Kowledge Argument. *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/sum2010/entries/qualia-knowledge/>

O'Connor, T. (1994). Emergent properties. *American Philosophical Quarterly, 31*(2), 91-104.

O'Connor, T. (2000). *Persons and Causes*. Oxford: Oxford University Press.

O'Connor, T., & Wong, H. Y. (2005). The metaphysics of emergence. *Nous, 39*(4), 658-678.

O'Connor, T., & Wong, H. Y. (2012). Emergent Properties. *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/spr2012/entries/properties-emergent/>

O'Dea, J. (2002). The indexical nature of sensory concepts. *Philosophical Papers, 31* (2), 169-181.

O'Regan, K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 883-917.

Oppenheim, P., & Putnam, H. (1958). The unity of science as a working hypothesis. In H. Feigl & G. Maxwell (Eds.), *Concepts, Theories, and the Mind-Body Problem* (pp. 3-36). Minneapolis: University of Minnesota Press.

Owens, D. (1992). *Causes and Coincidences*. Cambridge: Cambridge University Press.

Papineau, D. (2000). The Rise of Physicalism. In M. Stone & J. Wolff (Eds.), *The Proper Ambition of Science* (pp. 174-208). London: Routledge.

Papineau, D. (2002). *Thinking about consciousness*. Oxford: Clarendon Press.

Papineau, D. (2007). Phenomenal and Perceptual Concepts. In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 111-144). Oxford: Oxford University Press.

Papineau, D. (2008). Must a Physicalist be a Microphysicalist. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation and Causation* (pp. 126-148). Oxford: Oxford University Press.

Papineau, D. (2010). What Exactly is the Explanatory Gap? *Philosophia, 39*(1), 5-19.

Paul, L. A. (2007). Constitutive Overdetermination. In J. K. Campbell (Ed.), *Topics in Contemporary Philosophy vol. 4: Causation and Explanation*. Cambridge: MIT Press.

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.

Pettit, P. (1993). A Definition of Physicalism. *Analysis, 53*(4), 213-223.

Pettit, P. (1994). Microphysicalism without Contingent Micro-Macro Laws. *Analysis, 54*(4), 253-257.

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

Polger, T. (2008). H2O, 'Water', and Transparent Reduction. *Erkenntnis, 69*(1), 109-130.

Prinz, J. (2000). A Neurofunctional Theory of Visual Consciousness. *Consciousness and Cognition, 9*, 243-259.

Prinz, J. (2002). *Furnishing the mind*. Cambridge: MIT Press.

Prinz, J. (2007a). Mental Pointing: Phenomenal Knowledge without Concepts. *Journal of Consciousness Studies, 14, 9*(10), 184-211.

Prinz, J. (2007b). The Intermediate Level Theory of Consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciousness* (pp. 247-260). London: Blackwell.

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37-48). Pittsburgh: University of Pittsburgh Press.

Pylyshyn, Z. (2002). Mental Imagery: in search of a theory. *Behavioral and Brain Sciences, 25*(2), 157-237.

Pylyshyn, Z. (2003). *Seeing and visualizing: it's not what you think*. Cambridge, Mass.: MIT Press, Bradford Books.

Raatikainen, P. (2010). Causation Exclusion and the Special Sciences. *Erkenntnis, 73* (3), 349-363.

Ramsey, W. (2011). Eliminative Materialism. *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/spr2011/entries/materialism-eliminative/>

Rorty, R. (1965). Mind-Body Identity, Privacy, and Categories. *Review of Metaphysics, 19*, 24-54.

Russ, J. C. (2007). *The Image Processing Handbook* (5th ed.). New York: CRC Press.

Russell, B. (1927). *The Analysis of Matter*. London: Kegan Paul.

Savitt, S. (1974). Rorty's Disappearance Theory. *Philosophical Studies, 28*, 433-436.

Scerri, E. (2007). Reduction and Emergence in Chemistry-Two Recent Approaches. *Philosophy of Science, 74*, 920-931.

Schaffer, J. (2009). On What Grounds What. In D. Chalmers, D. Manley & R. Wasserman (Eds.), *Metametaphysics* (pp. 347-383). Oxford: Oxford University Press.

Schaffner, K. (1969). The Watson-Crick Model and Reductionism. *British Journal for the Philosophy of Science, 20*, 325-348.

Schiller, P. H., & Carvey, C. E. (2005). The Hermann grid illusion revisited. *Perception, 34*(11), 1375-1397.

Schwitzgebel, E. (2010). Introspection. *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/fall2010/entries/introspection/>

Seguin, E. G. (1886). A contribution to the pathology of hemianopsis of central origin (cortex hemianopsia). *Journal of Nervous and Mental Diseases, 13*, 1-38.

Shoemaker, S. (1980). Causality and Properties. In P. Inwagen (Ed.), *Time and Cause* (pp. 109–135). Dordrecht: D. Reidel.

Shoemaker, S. (1998). Causal and Metaphysical Necessity. *Pacific Philosophical Quarterly, 79*, 59–77.

Shoemaker, S. (2003). Realization, Micro-realization, and Coincidence. *Philosophy and Phenomenological Research, 67*, 1-23.

Shoemaker, S. (2007). *Physical Realization*. Oxford: Oxford University Press.

Shrader, W. (2009). Shoemaker on emergence. *Philosophical Studies*, forthcoming.

Silberstein, M., & McGeever, J. (1999). The Search for Ontological Emergence. *Philosophical Quarterly, 49*, 182-200.

Smart, J. (1959). Sensations and Brain Processes. *Philosophical Review, 68*, 141-156.

Smart, J. (1980). Physicalism and emergence. *Neuroscience, 6*, 109-113.

Smart, J. (2008). The Identity Theory of Mind. *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/fall2008/entries/mind-identity/>

Smith, A. D. (2010). Disjunctivism and Illusion. *Philosophy and Phenomenological Research, 80*(2), 384-410.

Snowdon, P. (2010). On the what-it-is-like-ness of experience. *The Southern Journal of Philosophy, 48*(1), 8-27.

Speaks, J. (2011). Theories of Meaning. *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/sum2011/entries/meaning/>

Spurrett, D. (1999). *The Completeness of Physics.* University of Natal, Durban.

Spurrett, D., & Papineau, D. (1999). A note on the completeness of 'physics'. *Analysis, 59*, 25-29.

Stoljar, D. (2000). Physicalism and the Necessary *A Posteriori*. *Journal of Philosophy, 97*, 33-54.

Stoljar, D. (2001). Two conceptions of the physical. *Philosophy and Phenomenological Research, 62*, 253-281.

Stoljar, D. (2005). Physicalism and Phenomenal Concepts. *Mind and Language, 20*, 469-494.

Stoljar, D. (2009a). Physicalism. *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, E. N. Zalta (Ed.), from <http://plato.stanford.edu/archives/fall2009/entries/physicalism/>

Stoljar, D. (2009b). The Argument from Revelation. In R. Nola & D. Braddon-Mitchell (Eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 113-138). Cambridge: MIT Press.

Sturgeon, S. (1998). Physicalism and Overdetermination. *Mind, 107*(426), 411-432.

Swoyer, C. (1982). The Nature of Natural Laws. *Australasian Journal of Philosophy, 60*, 203-223.

Thompson, E. (2000). Comparative Color Vision: Quality Space and Visual Ecology. In S. Davis (Ed.), *Color Perception: Philosophical, Psychological, Artistic and Computational Perspectives* (pp. 163-186). New York: Oxford University Press.

Titchner, E. (1929). *Systematic Psychology: Prolegomena*. New York: MacMillan.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(42), 1-22.

Turin, L. (2006). *The Secret of Scent: Adventures in Perfume and the Science of Smell*. London: Ecco.

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

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

Tye, M. (2003). A theory of phenomenal concepts. In A. O'Hear (Ed.), *Minds and Persons* (pp. 91-105). Cambridge: Cambridge University Press.

Tye, M. (2007). Philosophical Problems of Consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciouness* (pp. 23-35). Oxford: Blackwell.

Tye, M. (2009). *Consciousness Revisited: Materialism without Phenomenal Concepts*. Cambridge: MIT Press.

van Cleve, J. (1990). Mind-Dust or Magic? Panpsychism Versus Emergence. *Philosophical Perspectives, 4*, 215-226.

Van Gulick, R. (2001). Reduction, emergence and other recent options on the mind/body problem: A philosophic overview. *Journal of Consciousness Studies, 8* (9-10), 1-34.

Van Gulick, R. (2011). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*: URL = <http://plato.stanford.edu/archives/sum2011/entries/consciousness/>.

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science, 224*, 1260-1262.

Wallis, G., & Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences, 3*(1), 22-31.

Weiskrantz, L. (1997). *Consciousness lost and found*. Oxford: Oxford University Press.

Wertheimer, M. (1958). Principles of perceptual organization. In D. C. Beardslee & M. Wertheimer (Eds.), *Readings in perception* (pp. 115-135). Princeton: Van Nostrand.

White, S. L. (2007). Property Dualism, Phenomenal Concepts, and the Semantic Premise. In T. Alter & S. Walter (Eds.), *Phenomenal Knowledge and Phenomenal Concepts* (pp. 210-248). Oxford: Oxford University Press.

Williamson, T. (2003). Blind reasoning: understanding and inference. *Proceedings of the Aristotelian Society, 77*, 249-293.

Wilson, J. (2005). Supervenience-based Formulation of Physicalism. *Nous, 39*, 426-459.

Wilson, J. (2006). On characterizing the physical. *Philosophical Studies, 131*(1), 61-99.

Wilson, J. (2010). Non-reductive Physicalism and Degrees of Freedom. *The British Journal for the Philosophy of Science, 61*, 279-311.

Witmer, G. (2001). Sufficiency Claims and Physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and Its Discontents* (pp. 57-73). Cambridge: Cambridge University Press.

Woodward, J. (1997). Explanation, Invariance, and Intervention. *Philosophy of Science, 64*, S26-S41.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Worley, S. (2006). Physicalism and the Via Negativa. *Philosophical Studies, 131*(1), 101-126.

Yablo, S. (1992). Mental Causation. *The Philosophical Review, 101*, 245-280.

Yates, D. (2009). Emergence, downwards causation and the completeness of physics. *Philosophical Quarterly, 59*(234), 110-131.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308.

Zeki, S. (1983). Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelength and colour. *Neuroscience, 9*(741-756).

Zöllner, F. (1860). Ueber eine neue Art von Pseudoskopie und ihre Beziehungen zu den von Plateau und Oppel beschrieben Bewegungsphaenomenen. *Annalen der Physik, 186*(7), 500-525.