The World From Within: An Investigation into the Hard Problem of Consciousness from the Perspective of Bayesian Cognitive Science

Peter Kuhn

Doctoral Thesis, Philosophy

Goethe Universität Frankfurt

Advisor: Prof. Dr. Wolfgang Barz

Second Advisor: Prof. Dr. André Fuhrmann

February 13, 2024

Abstract

This thesis develops a naturalist theory of phenomenal consciousness. In a first step, it is argued on phenomenological grounds that consciousness is a representational state and that explaining consciousness requires a study of the brain's representational capacities. In a second step, Bayesian cognitive science and predictive processing are introduced as the most promising attempts to understand mental representation to date. Finally, in a third step, the thesis argues that the so-called "hard problem of consciousness" can be resolved if one adopts a form of metaphysical anti-realism that can be motivated in terms of core principles of Bayesian cognitive science.

Acknowledgements

Special thanks to my supervisor Prof. Wolfgang Barz for support, criticism and the freedom to develop my ideas. I would also like to thank Prof. André Fuhrmann, both for his willingness to serve as second advisor as well as for the opportunity to pressure test some of the contained ideas in the context of his Forschungskolloqium. Further I have to thank Henning Lütje, Daniel Weger, Elisabeth Wazcek, Max Hellriegel-Holderbaum and Jannik Luboeinski for volunteering to proofread various parts of the text at various stages of writing, as well as offering helpful advice. Also, I want to thank Tim König, Phillip Hey, Karl Friston, Maxwell Ramstead, Wolfgang Schwarz and David Chalmers as well as five anonymous referees assigned to the two paper projects that have developed from the text for helpful feedback and/or discussion. Finally, I want to thank Anna. None of this would have been possible without her.

Contents

Introduction

Conscious experience has a strange dual character. On the one hand, it seems to be the very thing we are most intimately familiar with. But at the same time, we are still far from a consensus regarding how consciousness could fit into the natural world, and whether it can so fit at all. The task of this thesis will be to develop a philosophical account of consciousness's place in the natural world, inspired by contemporary theories from neuro- and cognitive science.

While current scientific ideas, particularly those of Bayesian cognitive science, will figure hugely in what follows, the account will be distinctly philosophical. That is because the problem of conscious experience is ultimately philosophical or conceptual in nature. The problem is not primarily that we don't understand how conscious states are realized by specific neuronal ones, but that it seems that we could not even properly understand such a claim in the first place. The issue is not that we do not understand how consciousness arises from brain states, but that we don't know how consciousness *could* so arise.

This intuition was best expressed in David Chalmers' distinction between the *easy* and the *hard problem of consciousness*. The easy problem refers to the problem of explaining the functional properties of conscious states. This involves how certain stimulations and internal processes of the brain cause conscious experiences and how conscious experiences in turn figure in the control of behavior. On the other hand, the hard problem is the problem of figuring out how brain states could explain conscious experience all.[1]

Maybe the best way of illustrating the hard problem is in terms of the thought experiment of the *philosophical zombie*. A philosophical zombie is a molecule by molecule replica of a conscious being that is in fact not conscious at all. Philosophical

[1] Chalmers 1995.

zombies seem to be possible. That is, it seems that there is no overt error implicit in the idea. But if zombies are possible, then it seems consciousness can't be wholly explained in terms of brain function because the occurrence of the brain function is compatible with the absence of any conscious experience.

There are two boundary conditions I will set for the following investigation into the nature of conscious experience. First, I will assume that a form of *naturalism* is correct. While I will take consciousness and the hard problem seriously, I will also assume that no appeal to special entities that fall outside the scope of science is necessary in order to understand consciousness. The relevant naturalism will be *methodological* in nature. A methodologically naturalist theory is one that does not make any central assumptions that aren't backed or motivated by current natural science. It will be a central insight of later chapters that methodological naturalism is incompatible with *metaphysical naturalism*, the view that natural science captures the ultimate nature of things.

All this boils down to is that if the main argument developed in this thesis is successful, at the end of the final chapter it should become clear why the fact that zombies are conceivable is does not entail that there must be some irreducible mind-stuff as part of the constituents of reality. Seeing this will require that we change the way we think about a number of philosophical issues, like the nature of representation, truth and the relation of explanation and fundamental ontology. But ultimately all those revisions will be firmly rooted in our methodological naturalism.

The second boundary condition we will adhere to is that *strong illusionism* is false and *weak illusionism* is true at most in a very limited sense. *Illusionism* is the claim that our central intuitions about experience are in fact wrong. *Weak illusionism* is roughly the claim we are wrong about some of the superficial characteristics of consciousness. A *strong illusionist* claims that consciousness in fact does not even exist.[2] In effect, the second boundary condition says that consciousness exists and that our naive intuitions about it are roughly correct. It will turn out that where philosophers who hold the hard problem to be insoluble in a naturalistic manner go astray is not in immediate introspective intuitions, but in how they interpret these intuitions philosophically. Thus, I will argue that consciousness really has

[2]Frankish 2016.

all the properties it appears to have in naive introspection, however it is easy to be lead astray when theorizing based on those introspective facts.

In sum, we will develop a *non-illusionist naturalistic account of consciousness*. This theory will be developed in three major steps. I will now give an overview of each of these.

Step One: Phenomenological Analysis

Part one will focus on a phenomenological analysis[3] of the conscious experience, thereby delineating the target of our investigation - consciousness - more clearly. While this approach may be superficially at odds with our methodological naturalism, after all phenomenology is not an exact natural science, I take it as evident that the fact that consciousness seems problematic at all is grounded in our direct experience of it. There is no hard problem independently of our introspective intuitions. It is thus only natural to delineate conceptually, as precisely as possible, what exactly it is we are experiencing.

The central insight we will gather from this investigation will be that conscious experiences are a kind of representational state. Representational states are states that are essentially *about* something, they have an object. I will demonstrate that the best way of understanding consciousness is in terms of what conscious states are about. It follows that, if we could fully understand the nature of representational content in terms of brain states, this would amount to an understanding of the nature of consciousness.

A further insight of our phenomenological analysis will be that conscious representational states are characterized by the special kinds of content they represent. In particular, they represent special appearance properties like the intrinsic sense of blueness you experience when you see the cloudless sky. Thus, the distinctive challenge in understanding the nature of consciousness will be to understand how brain states come to represent these appearance properties.

Readers who aren't interested in the nitty-gritty may profit from merely skimming chapter two in part one and then jump straight to part two. You should still be able to follow my general argument.

[3] I regard my treatment of these subjects as part of what is sometimes called *analytical phenomenology*, not to be conflated with the phenomenological tradition going back to Husserl.

Step Two: Metasemantics

Building on the insights that conscious states are a certain kind of representational state, section two will investigate the prospects of accounting for the emergence of contentful states in terms of brain function. We will thus engage the question of the nature of meaning or metasemantics. Building on insight from the previous chapter, we will see that classical "tracking" theories of representational content suffer major difficulties and should be abandoned.

I will then give a very brief overview over the research paradigm of Bayesian cognitive science. Bayesian cognitive science conceives of the brain as engaged in a continuous attempt to predict future stimulation, thereby performing approximate Bayesian inference. Action and perception are thought to be emergent features of this underlying predictive activity. I will argue that we should conceive of representational mental states as being imbued with content in virtue of their causal role in this predictive process. This will result in a kind of inferential role semantics about mental states that explains representational content by a state's role in inference.

A central novel idea of this part of the project is that we can conceive of the appearance properties that are distinctive of the content of consciousness as the content of specific low-level representations of sensory input. These states have an inferential role that makes them represent appearance properties. Thus, I will argue that Bayesian cognitive science, in combination with an inferentialist metasemantics, has a natural and rather intuitive way of accounting for the nature of conscious representational content.

Step Three: Realism and Consciousness

The third and final step of the argument will engage the metaphysical background assumptions that are implicit in the usual theorizing about consciousness. In particular, I will argue that the hard problem of consciousness is based on a fallacious doctrine of metaphysical realism. According to metaphysical realism there is some wholly mind-independent way reality is and science and philosophy strive to capture this reality. Hilary Putnam illustrated this metaphysical prejudice with the idea of a God's eye point of view: For the metaphysical realists there is some way reality

really is, independently of how it appears to empirical observers, even under ideal epistemic conditions.[4] Against metaphysical realism I will argue for what I call model-relative realism. Inspired by Bayesian cognitive science, it makes sense to think of the true state of reality as relative to an organism's biologically determined mental makeup. Reality is co-constituted by the mental structures evolution has created *together with* the environment an organism is embedded in. Trying to pull these poles apart results in incoherent abstractions.

This has deep implications for the nature of consciousness for it entails that the very presuppositions that constitute the hard problem may be mistaken. When modern metaphysicians think about consciousness they do so precisely from a God's eye point of view, or so I will argue. From the alternative metaphysically anti-realist perspective it makes sense to hold that consciousness is neither physical nor irreducibly mental. The felt dichotomy turns out to be the result of a confused way of thinking about fundamental ontology.

Obviously, the project of this thesis is hugely ambitious, so ambitious indeed that it is virtually doomed to failure from the start. I just have only one excuse for my approach: Over the years I have come to the conviction that the hard problem of consciousness can only be satisfactorily solved if we revise some our central philosophical assumptions. I will suggest one such possible revision. May others do a better job.

[4] Putnam 1977.

Step One:

Phenomenological Analysis

1 Representationalism

In order to make sense of the complexity of the mental realm philosophers have drawn a number of useful distinctions between different aspects of mentality. Maybe the most influential of these distinctions is that between phenomenal consciousness and representational content. In this chapter we will study the relation of these realms. Our fundamental insight will be that the former is included in the latter: Phenomenal consciousness is nothing but a specific kind of mental representation.

Phenomenal consciousness is our central explanatory target. It refers to the qualitative dimension of experience. The best characterization of this phenomenon is Thomas Nagel's contention that if some being possesses phenomenal consciousness, then "there is something it is like" to be that very being.[1] I will also call this what-it-is-likeness of an experience its *phenomenal character*.

Phenomenal character describes a maximally specific characterization of a state of consciousness. A *phenomenal property* on the other hand is a less specific attribute of such a state. The former captures the total quality of your experience at a particular point in time while the latter picks out a specific experiential quality like seeing something grey, tasting hot coffee or hearing the pattering rain.

These are loose definitions that presuppose in some sense that the reader already knows what I am talking about. Presumably this flaw cannot be alleviated. It is an inconvenience we have to deal with. Our inability to offer good definitions for some domain is no reason to suppose that the phenomena within this domain are any less worthy of study, or any less in need of explanation. Contrary to the school of philosophy that insists on clear definitions, it seems that such definitions typically emerge *after* the interesting conceptual and empirical work is done. In fact, the theory

[1] Nagel 1974, p. 436.

I will expound in this thesis will partly justify the intuition that consciousness cannot be exhaustively defined. To know consciousness fully you have to experience it.

Conceptually speaking we have a much tighter handle on the second aspect of the mental, its *representational content*. While there are many conflicting ways to characterize this phenomenon, I will say that a state is representational if it is *about* some distinct state of affairs. More precisely, representational states are associated with *conditions of satisfaction*, conditions relative to which we can classify a state as satisfied or unsatisfied. Colloquially, we can refer to these conditions as the *content* that a representational state has or represents.[2]

I left the notion of satisfaction unspecified on purpose. Different representational states are associated with different notions of satisfaction. A belief can be *true* or *false*, a desire can be *fulfilled* or *unfulfilled*, a perception *veridical* or *non-veridical*. Formally however, all these can be expressed as functions from possible scenarios to (continuous or discontinuous) values of satisfaction. Again, I will talk of *representational properties* to describe the most general kind of features of representational states *qua* representational state. Therefore, being about Paris, believing the neighbour's cat to be grey and having a content at all, are all representational properties.

It is also common to associate representational states with *intentional objects*. I will understand these rather loosely as the objects to which the conditions of satisfaction of the state pertain. If I believe that the cat is on the mat, then the cat, and maybe the mat, are the intentional objects of this state. Note that the object of a representational state can be an *ostensible* object that need not exist. Hallucinations have intentional objects, too. We can understand conditions of satisfaction in a roughly descriptive manner that can be expressed as "There exists an object that is such and such." If there is such an object, the content will be satisfied. If there isn't, then it will not be satisfied. Note that the object-talk may sometimes be misleading. What is *the* object of experience when I am looking at a forest? Are there many objects (the trees) or is there one (the forest)? It seems you can cut this cake any way you like. In my view talk about intentional objects is just a convenient way of talking about representational content.

[2]To say that a state *represents a content* is in some sense a category mistake. The state represents an object or a state of affairs and this fact constitutes its content. As long as this fact is kept in mind however, the formulation is harmless and canonically used in the literature, see for instance Chalmers 2003b.

A crucial way of differentiating between different families of representational states is the notion of a *direction of fit*. Directions of fit specify the manner in which the content of a representational state ought to be satisfied. I will speak of a *mind-world* direction of fit, if the satisfaction of the content is up to the mind. This is true for beliefs and perceptual states, for instance. The states purport to capture the world as it is. If the world is different from what you believe or perceive, this is *your fault*, not a fault of the world.

Desires, wishes and intentions on the other hand have a *world-mind* direction of fit. These states don't merely describe the world, they oblige to change it. A failure of fit is, so to speak, a flaw of the world, not of the desire itself.[3] The direction of fit can be conceptualized as a note on who is making a mistake if content and world come apart. For states with a *mind*-world direction of fit the mistake is on part of the mind, for states with a *world*-mind direction of fit the mistake is on part of the world. In this chapter we will focus on states with a mind-world direction of fit. This approach will only be justified when we later discuss Bayesian cognitive science and predictive processing and see how these kinds of states form an underlying unity.

Representational content can both be a feature of *occurrent* states or of *standing* states of agents. Perceptual states fall in the former category, beliefs in the latter. Perceptual states occur at a particular point in time (Seeing that there is a pile of books on my desk), beliefs are relatively stable over long periods (your belief that Paris is the capital of France). Perceptual scenarios are events, beliefs are typically held to be dispositional states of agents. Phenomenal consciousness, on the other hand, can either be attributed to agents (are bats conscious?) or their states (is bat echolocation conscious?). I will assume that agent consciousness is explicable in terms of states of consciousness, i.e. saying that the bat is conscious just means that at least some state of the bat is conscious.

We will now get to the central thesis of this chapter, namely that phenomenal properties are a particular kind of representational properties. Consciousness, I will argue, *is* a form of relatedness to an intentional object.

[3] For a classical treatment of representational content in terms of conditions of satisfactions and directions of fit, see Searle 1983.

1.1 The Varieties of Representationalism

The thesis that phenomenal properties are identical to representational properties is often called *intentionalism* or *representationalism*. I will stick to the latter term. In this section I will elaborate the thesis and get some misconceptions out of the way.[4]

Prima facie, representationalism shouldn't be a too surprising position. Phenomenal consciousness is the way a certain state is like for an agent. Nagel famously introduced the term by contemplating what it is like to be a bat. The phenomenal character of the bat's experience at a certain point in time might arguably also be characterized as *the way the world appears to the bat* and it is very plausible that an appearance involves a representational state in the sense that appearances can be accurate or misleading. Appearances seem to have conditions of satisfaction with a mind-world direction of fit. We will later see how to make this suspicion precise.

Enemies of representationalism typically hold that, while phenomenal consciousness might have representational aspects, the two dimensions of the mind still have to be kept firmly apart. On this view, the phenomenal properties of an experience might best be thought of as consisting of a kind of *mental paint*[5] analogous to the paint used to paint a canvas. The resulting picture has representational properties, but the properties of the paint covering the canvas aren't exhausted by how it depicts its object. In the same manner, defenders of mental paint typically hold that the phenomenal properties enter into the representational properties of experiences, but phenomenal character is not exhausted by the representational aspects.[6]

The idea of mental paint *prima facie* makes phenomenological sense and there are a number of experiences that make the view compelling. The phenomenology of pain and joy, for instance, hardly seems to be exhausted by the manner in which these experiences depict the world. The phenomenology of a dark mood

[4]There is an alternative formulation of representationalism that claims that *phenomenal character is representational content* (see for instance Tye 2008, p.112.) To my mind, the most obvious interpretation of this sentence yields nonsense. Phenomenal character is most naturally conceived as a kind of complex phenomenal property. Representational content however, arguably is an abstract object and thus both cannot be identical. There is of course the charitable interpretation that phenomenal character is identical to *the property of having* a certain content. But this just yields a confusing way of stating the simpler version, namely that representationalism is the thesis that phenomenal properties are representational properties.

[5]This term is used in Harman 1990 as a slur and taken up as a self-attribution in Block 1996.

[6]There is also the position of separatism that holds that phenomenal consciousness and representational content are utterly independent of one another. This seems highly implausible. At any rate, the argument given below will also apply to separatism.

does not seem to be exhausted by the fact that the grass appears less green. We will discuss such hard cases in chapter four and see why mental paint really offers no explanatory surplus in dealing with them. For now it will be helpful to focus on more or less uncontroversial perceptual cases.

There is an ongoing debate about how far down the roots of phenomenal consciousness reach. While representationalists are committed to the claim that all phenomenally conscious states are representational states, the reverse does not hold true. In particular, one might hold that thought and conceptually structured cognition are associated with a characteristic phenomenology, or one may deny this. Defenders of such *cognitive phenomenology* hold that there can be something it is like to think that 6 times 13 equals 78, for instance, while its enemies hold that this is not generally the case.

The issue is delicate, and it seems hard to find decisive reasons either way. My own intuition tells me that there are abundant examples of cognitive phenomenology, but I find it hard to formulate a conclusive argument. The best I can come up with is the strong intuition that *phenomenal duplication*, i.e. creating an entity that shares my phenomenal properties, necessarily entails the duplication of some thought content. I cannot conceive an agent that is phenomenally identical to me but thinks wholly different thoughts.[7] But this is still a brute intuition. This is why I will try to keep the following phenomenological discussion neutral on this question and comment on how my view is compatible with cognitive phenomenology where I deem it necessary. Later we will see that, according to a plausible construal of Bayesian cognitive science, the limits between the perceptual and thought are fleeting. This may explain why cognitive phenomenology is hard to pin down introspectively: There just may be no clear point where the perceptual phenomenology ends and thought begins.

Representationalism isn't committed to the position that all there is to phenomenal consciousness is representational content or that all we have to do to understand consciousness is to study how brain states get associated with conditions of satisfaction. According to representationalism, having a certain representational content is a *necessary* feature of phenomenally conscious states, not a *sufficient*

[7] Horgan and Tienson 2002.

one. This is because two different representational states can not only differ in content, but also in *how* that content is represented.

Let us call representational properties that merely pertain to the content of the relevant state without pertaining to the way it is represented *pure representational properties*. Then *pure representationalism* is the view that phenomenal properties are identical to pure representational properties. Pure representationalism holds that consciousness is a matter of representational content only. On this view, an explanation of representational content yields an explanation of conscious experience.

Pure representationalism is widely deemed implausible because it seems that content alone is not sufficient to fix phenomenal character. There arguably are a variety of different kinds of unconscious representational states. Cognitive science postulates unconscious sub-personal representations, psychodynamic theories postulate unconscious desires, wishes and thoughts and even folk-psychology postulates standing states like beliefs and desires that by themselves are unconscious. Therefore, it is argued, there must be something other than content that results in a difference between conscious and unconscious mental representation.

A similar problem for pure representationalism may arise from inter-modal differences. I may see that it is starting to rain, and I may hear that it is starting to rain. I may see that the rain pounds against the window, of I may hear or even feel it. On some views of the content of perception these might be expressed as phenomenally different perceptual states that carry the same representational content. If this is correct, then pure representationalism must be false.

This is why most representationalists defend *impure representationalism*, the view that phenomenal properties are identical to impure representational properties. *Impure representational properties* characterize not only the content of a state, but also the manner in which this content is represented. Therefore, conscious and unconscious representational states are taken to differ not in content, but in the way these contents are represented[8]. Some claim that the same is true for inter-modal differences.[9]

There is no agreement among impure representationalists on the supposed manner of representation that turns unconscious into conscious representational content or an auditory experience into a visual one. There are a variety of options. Many early

[8]Jackson 2003; Chalmers 2004; Seager and Bourget 2017; Smithies 2019.
[9]This route is taken in Lycan 1996 and Egan 2006.

defenders of representationalism proposed that conscious representational states have to have certain functional characteristics, like being poised for usage by a concept using system.[10] Dualists that are prone to representationalism on the other hand, may claim that the right manner of representation is somehow irreducibly phenomenal. On these views phenomenal properties involve a content being represented *consciously*.[11]

It might be suspected that impurity somehow makes representationalism explanatorily worthless. But this is not the case. All it entails is that, if it is true, explaining representational content does not amount to explaining phenomenal character. Still, according to impure representationalism, explaining representational content is *necessary* to explain consciousness. Therefore, the truth of any kind of representationalism would rule out theories that take representational content to be ontologically independent of conscious experience.

To demonstrate that even impure representationalism involves a substantial claim about the nature of experience, let us consider a popular account of consciousness that is incompatible with it. *Integrated information theory* holds that there is some quantity Φ, that roughly measures the recursion of causal connections within a complex system, that is proportional to the level of consciousness of a relevant system. What is interesting for our purposes is that a high level of Φ seems to be independent of any representational capacities of the system.[12] Integrated information and representational properties are supposed to vary independently. Integrated information is thus mental paint *par excellence*. If you think that the following argument for representationalism is convincing then this also offers a decisive reason to reject integrated information theory.[13] Insofar as the theory is one of the main

[10]The formulation can be found in Tye 1995, similar accounts are given in Dretske 1995, Lycan 1996 and Jackson 2003.

[11]Chalmers 2004. A similar view is also suggested without conclusively being embraced in the latter parts of Kriegel 2017.

[12]There are two ways of seeing this. First, we can plausibly assume some kind of functionalism about representational properties, i.e. that representational properties of some state are determined by its long-armed or short-armed (Harman 1987) causal role. Integrated information theory typically denies functionalism (Tononi et al. 2016). Thus, integrated information theory denies representationalism. Secondly, integrated information theory is typically motivated based on certain axioms about the nature of experience that purport to apply to experience, rather than its object. As an aside, from the representationalist vantage point it is plausible that these axioms are based on a mis-attribution of properties of the object of experience to the experience itself.

[13]Or at least a reason to significantly alter the formulation. Maybe it might be claimed that integrated information is somehow intrinsically representational. However, considering the way the theory is standardly motivated, this seems *prima facie* implausible, see previous footnote.

contenders for a scientific account of consciousness, the question of representationalism is central not only to philosophy but also to the science of consciousness.

This section served the purpose of giving the reader a brief overview over the idea of representationalism. I have elaborated the difference between pure and impure variants of the theory. The following section will develop an argument for representationalism.

1.2 The Transparency Argument

My main argument for representationalism will rest on the so-called *transparency* of experience. Intuitively, when we try to grasp the nature of our conscious states we glide through them to the states of the world. I try to contemplate the phenomenal character of the experience of the blue sky and I end up contemplating the appearance *of the sky itself*. This transparency, I will argue, indicates that the nature of experience, as it appears in introspection, is exhausted by its representational characteristics. Phenomenal character is nothing more than a specific kind of aboutness.

To me, the datum of transparency seems clear enough for most experiences. I take a sip of coffee, and it is the coffee that appears hot and bitter, not the experience. I see fog hanging over the nearby hills, and it is the fog that is grey, not my seeing. And even when I hear birds singing, then it is the singing that appears beautiful, not the hearing. If all this is true, and if it is universally true of all experiences, then representationalism seems to follow naturally. The qualitative feel of an experience isn't some property that directly pertains to the experience itself, rather it pertains to its objects. What pertains to the object of experience must be part of its representational content.

Expressed another way, transparency indicates that having a conscious experience is not like looking at a mental picture of something. If this were so I could meaningfully differentiate introspectively between the properties of the (mental) paint and the properties of the depicted object. Transparency, if phenomenologically valid, seems to show that I cannot differentiate the two.

Before I embark on explicating the argument in detail I want to get a possible point of confusion out of the way. Debates about transparency roughly concern issues of *introspection*, i.e. our capacity to know our own minds. However, introspection may

also be understood in a more demanding way that is closer to its etymological roots, namely as a capacity to *see inside* and inspect our own mental states. If introspection is taken in the second sense however, then representationalism precludes the very possibility of introspection. If consciousness *is* a form of directedness at the world, then it has no inside to see. I will understand the term introspection in the less ambitious sense as whatever means we are using when we attend to our own mental states. On this understanding, it is not a contradiction to say that introspection tells us that experience is transparent and thus representational.

The argument from transparency is usually attributed to G. E. Moore in the following often quoted passage.

> [T]hat which makes the sensation of blue a mental fact seems to escape us; it seems, if I may use a metaphor, to be transparent — we look through it and see nothing but the blue... the moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous.[14]

A much earlier version of this kind of reasoning appears in the writings of Charles Sanders Peirce. The fourth question in his *Questions concerning certain Faculties claimed for Man* is "Whether we have any power of introspection, or whether our whole knowledge of the internal world is derived from the observation of external facts?" (note that Peirce uses the term "introspection" in the more restricted sense)[15]. And, as with all questions in the essay, Peirce will deny this one. Rather,

> [...] we may derive knowledge of the mind from considerations of this sensation, but that knowledge would, in fact, be an inference from [...] a predicate of something external.[16]

As Moore is one of the most important popularizers and interpreters of Peirce, it seems highly probable, that Moore's argument was inspired by the latter.

[14]Moore 1903, p. 446.
[15]Peirce 1868.
[16]ibid.

While there are variants of the transparency argument scattered in the philosophical literature[17], the argument was reintroduced into the contemporary discourse by Gilbert Harman[18] and taken up by the two paragons of representationalism, Michael Tye[19] and Fred Dretske[20]. Harman writes:

> Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree "from here."[21]

When contrasted with the popularity of the argument from transparency in favour of representationalism, the lack of agreement on what the argument actually *is*, is interesting. My suspicion is that this is partly due to the formulation of the argument as a claim about *attention*. Going back to Moore's lucid description, transparency has been taken as the fact that if one *attends* to one's experience one ends up *attending* to the features of the object of experience. But of course, the fact that attending to A leads one to attend to B does not entail that A really can be analyzed in terms of B![22][23]

I will call the claim that in attending to properties of experience one inevitably ends up attending to the objects of experience *Moorean transparency*. Now Moorean transparency, I contend, is based on the more fundamental fact of what one may call *Peircean transparency*. In the above quote, Peirce doesn't make any special claim about attention at all, but rather suggests that the *properties* that we are aware of in experience are really *properties* of the objects of experience.[24]. If this is true, then there is a rather straight-forward formulation of the argument from transparency, namely:

> (1) All properties we are directly introspectively aware of are properties of the objects of experience. (Peircean Transparency.)

[17] See for instance Sartre 1982. Thomasson 2003 argues that Husserl can be interpreted as a defender of the transparency argument.
[18] Harman 1990.
[19] Tye 1995; Tye 2014.
[20] Dretske 1995.
[21] Harman 1990, p.39.
[22] For a lengthy argument to this conclusion, see Stoljar 2004.
[23] This kind of formulation can be found in Tye 1995 p.30, Dretske 1995 and as we already saw, Harman 1990 (though Harman seems to realize that Moorean transparency is based in Peircean transparency, see below).
[24] Arguments from Pieciean transparency may be found in Kriegel 2009, p. 68-71. Similar remarks can be found in Shoemaker 1994a

(2) If all properties we are directly introspectively aware of are properties of the objects of experience, then representationalism is true. (Peircean transparency implies representationalism.)

(3) Representationalism is true.

I already tried to motivate the first premise. At least in cases of perceptual experience, we are confronted with the properties of the objects of experiences, not directly with the properties of the experiences themselves. When I see the grey sky, it is the sky that appears grey, not the experience. Coming to know the properties of experience, as Peirce says, involves an *inference*. In chapter eight we will have much more to say on what kind of inference this is. The fact that I have an experience of a particular tone of grey is actually inferred from facts about the intentional object of my experience, the sky.

Note that insofar as the objects of experience are intentional objects and intentional objects may merely be *ostensible* objects, the first premise cannot be evaded on the grounds of illusions. It may turn out that the object of experience, in some sense, really does not have the properties I experience it to have. The wall may appear red, but this may turn out to be the result of red illumination. And in some sense this is a direct refutation of premise one. But this of course is not the intended reading of 'objects of experience'. These are intentional objects, i.e. objects experience seem to be about. In this sense, the intentional object of an experience of a redly illuminated white wall may be a red wall.

In the perceptual case transparency seems to be clear enough. However, as mentioned, there are harder cases, like pain and joy, where the transparency intuition is less clear. When I hurt my knee, isn't it the experience that is hurtful, and not just the knee? On the face of it, this would refute the first premise. At the moment we do not yet have the conceptual resources to deal with this claim. I therefore postpone the discussion of such difficult cases to chapter three.

The second premise might be less intuitively clear than the first one. Its validity depends on the implicit assumption that all phenomenal properties can be uncovered using introspection, or at least on the assumption that if there are *hidden* phenomenal properties, meaning phenomenal properties that are not disclosed to introspection, they are representational properties. So lets start by asking whether this assumption

may be rejected. This seems hopeless. While one might coherently hold that, maybe due to limited capacity, we cannot be introspectively aware of all phenomenal properties of our experience at once, this gives us no reason to suppose that these hidden phenomenal properties are somehow unlike the phenomenal properties we are introspectively aware of. So the implicit generalization in premise two, from phenomenal properties we know of to all phenomenal properties, seems plausible.

Given that there are no hidden phenomenal properties the second premise is relatively straight-forward. One can argue for it by *reductio*. If representationalism were false, there is some phenomenal property that is not a representational property. This phenomenal property would then be an introspectively apprehendable property that, whatever it is a property of, would not be a property of the object of experience. But then the left side of the conditional cannot be true. Thus, the second premise must be correct.

An attack on the second premise might be launched from the vantage point of *projectivism*. This is the notion that directly perceived properties really pertain to some aspect of our experience but are represented as though they were an aspect of the world. On this view the properties we directly perceive are properties of the ostensible object of experience in virtue of also being properties of the experience.[25] However, I think projectivism is hard to motivate phenomenologically, given the first premise. If we are never directly confronted with sensible properties as properties *of experience*, why should we believe they are such properties? Paul Boghossian and David Velleman motivate projectivism primarily with recourse to counter-examples to representationalism.[26] We will discuss various counter-examples of this kind in the final chapter of part one and see why none of them work. So the supposed attack on premise two collapses into an attack against premise one. In order to phenomenologically motivate their view projectivists will have to deny transparency from the start. Thus, if our defence of the datum of transparency in chapter three is successful it will also reject projectivism.

Furthermore, projectivism has to claim that the common sense take on perceptual experiences confronts us with a bewildering form of category mistake, namely a confusion of the vehicle of representation with its object. Isn't the conclusion that

[25]Boghossian 1989.
[26]Ibid.

the tomato is red because the experience of the tomato is red as absurd as the conclusion that Boghossian is to be found in chapter one of my doctoral dissertation, just because his name is written here? How could such an egregious error be built into the very structure of the perceptual system?[27]

A similar attack on premise two may be launched from the vantage point of sense-datum theories. Here, the properties that seem to be mere properties of the object of experience are in reality special properties of mental entities. Just as in the case of projectivism this would undercut premise two. But just as in the case of projectivism, given Peircean transparency, sense-data theories will be very hard to motivate phenomenologically. In effect, as it is already admitted that the relevant properties *seem* to pertain to the objects of experience, a sense-data theorist will have to argue for her theory on wholly non-phenomenological grounds.

A further criticism of the argument might be fueled by issues with the phrase "directly introspectively aware". But, as I suggested in the case of phenomenal consciousness, the quest for a definition might be misguided here. That in experience we are directly aware of certain properties is a fact everyone is familiar with and that we can confidently presuppose.

Much confusion has been generated by assuming that the primary datum of transparency is Moorean transparency, i.e. a claim about attention. Daniel Stoljar has argued that one cannot infer Peircean transparency from Moorean transparency.[28] While I am neutral on the issue,[29] I disagree with Stoljar that Peircean transparency is itself any less plausible than Moorean transparency. It may be didactically helpful to point out Moorean transparency in order to see that Peircean transparency is plausible as well, but one can also directly observe Peircean transparency introspectively.

Even devout critics of the transparency argument seem to agree that Peircean transparency would be a plausible ground for inferring representationalism. Stoljar holds that our introspection can only show that that Moorean transparency is true and thus he thinks we can reject the argument. And Amy Kind holds that Peircean transparency may be plausible for some experiences, but she holds

[27] See McGinn 1984, chapter 5.
[28] Stoljar 2004.
[29] I am not completely on board with the claim that transparency contradicts the claim that we know facts about experience by inference from the properties of objects (Stoljar 2004, p. 372).

it to be implausible as a universal thesis.[30] In summary, it seems there aren't very many philosophers who would object to the inference from transparency to representationalism. What we will have to discuss in detail in chapter three is whether Peircean transparency is indeed universal.

1.3 Summary

We have introduced the conceptual framework for thinking about the relation of phenomenal consciousness and mental representation. Furthermore, we established that *if* experiences are transparent in a Peircean sense, that is if the properties we are directly aware of in experience are properties of the object of experience, then it is highly plausible that consciousness is nothing but a kind of mental representation.

This sets the groundwork for a naturalist explanation of consciousness. For if consciousness is nothing but mental representation, explaining consciousness will amount to explaining mental representation. And mental representations many philosophers hold, are indeed naturalistically explainable. But before we can engage such a task we will have to first deal with the question of what kind of mental representation precisely consciousness is. We will engage this task by asking what it is precisely that our experiences represent. As we will see this is far from straightforward.

[30] Kind 2003.

2 Appearance Properties

If phenomenally conscious states are representational states, what is their content? What are conscious experiences about? Let's call the representational content of a state that is necessitated by its phenomenal properties its *phenomenal content*. Note that this is not to say whether phenomenal properties are metaphysically more basic than phenomenal content. Many representationalists hold that the proper direction of explanation is the other way around: Representational properties explain phenomenal ones. Still these representationalists should hold that if some state has certain phenomenal properties it necessarily has some specific content, i.e. a phenomenal content.

It is furthermore helpful to differentiate between the content a certain experience has by its very nature, its *intrinsic phenomenal content* and content it has contingently. For instance, if you are afraid of the dark, the color black may appear differently to you than if you are not. But the fear you associate with the dark is not part of the intrinsic phenomenal content of seeing blackness, but merely its *extrinsic phenomenal content*. It is part of your particular experience of blackness, not a part of experience of blackness as such.

Let's call the properties that our experiences intrinsically represent their objects to have *sensible properties*. The grey of the sky and the bitterness of the coffee *as they are represented in experience* are sensible properties. At some points I will refer to these properties as *s*-grey and *s*-bitter. They are the kind of properties that the ostensible objects would have to have for the intrinsic phenomenal content in question to be veridical.

Sensible properties aren't the same as what we might call *public properties*, the properties typically attributed by ordinary language predicates. Public and sensible properties regularly come apart and much confusion is generated this way. For instance, it may be that sensible colors are primitive intrinsic feels attributed to surfaces

in experience, while public colors are dispositions to reflect light of certain wavelengths. Having a blue experience and saying "This is blue" might thus involve very different conditions of satisfaction. I wont have much to say on public colors in this chapter.

This section investigates the nature of sensible properties. As it turns out, this is a less straightforward task than one might hope. I will refute three accounts according to which sensible properties are physical properties, primitive "edenic" properties, or modes of presentation. In the second section, I will introduce my positive account according to which sensible properties are ways of appearing.

2.1 The Riddle of the Senses

In this section I will reject three suggestions regarding the nature of sensible properties. These are first, the view that sensible properties are physical properties of the represented objects, secondly, the suggestion that they aren't properties at all, but modes of presentation of properties, and finally, that they are ostensible primitive properties of things external.

2.1.1 Are Sensible Properties Physical?

On first approximation one might think that the properties that are represented by our perceptual states must be ordinary external world properties with which their occurrence is caused by or correlates with. This is the route taken by so-called *tracking representationalism*. Here, perceptual representational content is understood as some kind of (further qualified) co-variational relationship between internal indicators and external states of affairs. On this view, the fact that some neuron in my brain represents cats is grounded in the fact that the probability that there is a cat in front of me is higher, given that this neuron fires.[1]

On the tracking view, internal states can't help but represent ordinary properties of the external world. And isn't that also just the most obvious route to take? Clearly s-greyness is nothing else than the property of reflecting a certain mix of wavelengths of light. Clearly s-bitterness is nothing but a certain chemical

[1]Dretske 1995 argues that the probability has to be 1. Tye 1995 requires a stronger causal relation.

property of the coffee. Sensible properties are physical properties of the world, or so tracking representationalism wants to make us believe.

As intuitive as this view might seem at a first glance, it is hard to take it serious upon reflection. The color of the clouded sky doesn't appear in my experience as though it had some disposition to reflect light, it just appears *grey*. It is some unspeakable quality, a simple and non-analyzable *thusness* spread across the cloudy sky. My experience of drinking doesn't represent the coffee to have a certain chemical composition. It represents it to be *bitter*, as having an unspeakable feel to it.

Note that it arguably is not a valid reply for the tracking representationalist to say that sensible properties are physical properties that for some reason *do not appear like physical properties*. The current challenge is to explain the nature of sensible properties, i.e. ways things appear in experience. The logical structure of this problem does not allow for a wedge to be driven between what the representational content of our experiences appears to be and what it actually is. It is the very appearance that we want to understand.

2.1.2 Is Phenomenal Content Fregean?

So sensible properties aren't physical properties. Rather, it may seem upon refection, they are *ways in which we represent* physical properties such that an experience of some sensible property doesn't necessarily disclose the nature of the physical reality that underlies it. Ways of representing are ordinarily conceived in terms of Fregean *modes of presentation*.

This suggestion requires some explication. *Modes of presentation* serve the purpose of relating cognitive systems to object in the world. Contemplating the meaning of identity statements like "The morning star is the evening star.", Frege realized that conceiving of identity as a mere relation between an object and itself makes such statements vacuous. That the morning star is the evening star expresses something deeper than just that planet Venus is planet Venus.[2] Frege's solution is to conceive of the morning star and the evening star as Venus *under different modes of presentation*. One and the same object (Venus) can relate to a cognitive system in two different manners.

[2]Frege 1892.

Contents conceived as stand-ins for states of affairs are usually called *Russelian*. Contents conceived as mediated as through modes of presentation are usually called *Fregean*. The current suggestion is that the phenomenological implausibility of the view that sensible properties are physical can be mitigated by supposing that phenomenal contents are Fregean rather than Russelian. They are modes of representation of physical properties.

An intuitive motivation why one may hold that phenomenal content is Fregean can be gained by contemplating the *inverted spectrum* thought experiment. Nonvert and Invert are functionally isomorphic twins.[3] However, Invert has an inverted spectrum relative no Nonvert. Where Invert sees *s*-green, Nonvert sees *s*-red and so on.

Prima facie, the inverted spectrum thought experiment is a threat to representationalism. In some sense, Nonvert and Invert represent *the same thing* when they see a tomato. But their experiences are obviously quite different. On the face of it, this looks like a counter example to representationalism. A natural representationalist response may be the following: Nonvert and Invert represent the same physical properties, but under different modes of presentation. Sensible green is the mode of presentation under which Nonvert represents physically green things, sensible red is the mode of presentation under which Invert represents green things.[4]

Unfortunately, this suggestion is phenomenologically implausible. Instead of contemplating interpersonal differences, like between Nonvert and Invert, consider inner personal ones. You look at a green piece of cloth and watch as it slowly turns red. You can't describe this transition as a mere change in mode of presentation! The difference obviously pertains to *the properties* that are represented and not just to *how* they are represented.[5] Examples of this kind strongly suggest that the relevant contents are Russelian rather than Fregean.

[3] If you believe that functional duplicates with different conscious states of consciousness are inconceivable, for the current purposes it should suffice to imagine that Nonvert and Invert are just functionally quite similar, not necessarily identical. We will discuss issues of conceivability and possibility in chapter ten.

[4] For a position of this kind see Hilbert and Kalderon 2000. One may also read Chalmers 2004 as defending such a position.

[5] Speaks 2009; Speaks n.d.

2.1.3 Are Sensible Properties Edenic?

There is something right about the approach of conceiving of colors in terms of modes of presentation and our final account is partly inspired by it. But on close phenomenological inspection I think the following conclusions seems hard to avoid: First, phenomenal content is Russelian. Secondly, the sensible properties that are manifest in experience are irreducibly primitive properties. In the terminology of David Chalmers, I will call these hypothetical properties *edenic properties*.[6]

Chalmers illustrates this view by considering the logically possible but non-actual world Eden, where the world is actually filled with edenic properties. In Eden, the grass is *actually green* and does not just have the disposition to cause green experiences. In Eden, the sky does not just reflect a certain wavelength of light, but it is actually primitively blue. In Eden, everyone perceiving an object correctly will necessarily enjoy the same perceptual phenomenology. The *edenic view* holds that our phenomenology is such that we perceive the world as though still inhabiting Eden "before the fall".[7]

If this is correct, then tracking representationalism must be false. However, the representational content of experience comes to be, it can't be via tracking relations to external objects. The reason is simple. Scientific investigation tells us that there are no edenic properties anywhere in our environment. Therefore, our internal states hardly can covary with their occurrence and some kind covariation is a necessary condition of representation according to the tracking view.[8] If the edenic view is correct we will need some other account of representation.

So let's consider the view that our experience represents edenic properties. This account avoids the phenomenological implausibility of the tracking view and its Fregean variation, but it runs into two different issues. First, it is not entirely clear whether the view that experiences represent edenic properties captures the intrinsic phenomenal content of perceptual experience very well. In particular, the edenic contents are *thick* in the sense that attributing an edenic property to an object entails a lot of commitments about it. In opposition to this, I will argue that the intrinsic phenomenal content is arguably *thin*, that is, it does not involve any complex commitments.

[6] Chalmers 2010.
[7] Ibid.
[8] Chalmers 2010, Mendelovici 2018, chapter 3.

According to the edenic view, it is part of perceptual content that perceiving the object in *this* particular way is in fact the *only way* to correctly perceive it. Accordingly the assumption of *intersubjective uniqueness* is thought to be directly built into the very intrinsic phenomenal content of the experience of sensible properties. In a similar way, according to the edenic view, sensible properties are intrinsically *observer-independent*. When one sees the sky as edenically grey, then one sees it as having this property independently of whether one is looking at it.[9]

It can reasonably be doubted whether intrinsic phenomenal content of typical sensory experience is thick in this way. Consider the claim that sensible properties are intrinsically intersubjectively unique. Imagine a hermit that never had any contact with other human beings. When the hermit lets his gaze wander across the cloudy sky and sees it as *s*-grey, is it really plausible that this ought to move him to conclude that other agents would see the sky the same way he does, even if the thought of other agents never once crossed his mind? It seems it is much more plausible that the bare phenomenology does not carry any commitments about intersubjective perception, i.e. that the intrinsic phenomenology is non-committal or thin.

Something similar holds for the notion of observer-independence. Look out of the window. Close your eyes. Open them. Does the scenery look as if it looked that way when your eyes were closed? Some readers may intuitively say yes, others no. But now ask yourself whether it is reasonable to suppose that one group must be right, given that we once again can just claim that our perceptual phenomenology is thin, meaning that it, all by itself, does not commit us to either view. It seems plain to me that the disagreement is insoluble and should be abandoned as a misunderstanding about how deep phenomenology reaches. We will return to this point below.

A second problem for the edenic view is that if our experience represents the world as being full of primitive edenic properties, but these properties aren't instantiated anywhere, then our experience must be massively and systematically falsidical. In fact, if the view is correct, there just is no such thing as consciously perceiving the world as it really is. The nature of experiences themselves is deemed deceptive.

At first this may not seem like a deep problem. In recent decades and maybe centuries the view that our experience of the world is massively mediated by un-

[9] Chalmers 2010.

conscious constructive mechanisms has become kind of a popular sentiment and the view that colors are somehow less real than normal physical properties is around at least since Galileo drew his famous distinction between primary and secondary qualities. For the modern mind, the view that colors are not really *out there* has lost its shocking connotations.

The degree to which the view that the systematic illusion ostensibly at work in our perception is problematic depends on how far-reaching one supposes the illusion to be. On phenomenological reflection it becomes quite clear that edenic properties are not limited to the domain of colors or tastes, i.e. the domain that one might call secondary qualities. Rather *all* possible contents that can be the object of perceptual consciousness seem to involve unanalyzable sensible properties and thus, according to the view under consideration, edenic properties.

A prime example would be the experience of space. It seems quite plausible that there must be an edenic property associated with the experience of space itself. This intuition is hard to communicate, however one might come closer to appreciating it by contemplating the fact that our experience of space is often associated with *visual* connotations. Euclidean space is nothing but a certain ordered set of points, i.e. there is really nothing *visual* about it. This seems to indicate that our ordinary, visually tainted experience of space is attributing edenic properties to it.[10] If space is devoid of edenic properties and our experience represents it as having such properties, our very experience of spatial relations must in some manner be illusory. This, on the face of it, seems even more implausible than the mere contention that colors are illusory. The more inclusive the systematic illusion is held to be the more one has to wonder whether the assessment of illusoriness is an artifact of our thinking about the nature of experience rather than a feature of naive perception.

For defenders of cognitive phenomenology, the mystery even deepens. If conceptual thought attributes edenic properties to its objects, it must be massively and systematically falsidical as well. This is especially troubling for those philosophers who take the experiential side of cognition to do important work in content determination like grounding the distinction between using the symbol "+" to denote addition

[10]This view is elaborated at great length in Thau 2002, chapter 5.

rather than some other function[11], or determining linguistic meaning[12]. How is cognitive phenomenology supposed to ground representation if it is systematically misrepresenting? A defender of the edenic view may bite the bullet of widespread illusion. But given that the view also seems to misdescribe the thin phenomenology of experience, we are certainly justified in searching for an alternative. All this already points in the direction that a suitably 'thinned out' edenic view may do the trick. This will be the route followed below.

So we are faced with a dilemma. Sensible properties are clearly properties in the full sense of the term, not just modes of presentation. But they can't be physical because they don't appear to be. And they can't be edenic because the edenic view seems to over-rationalize experience and make it illusory to an implausible degree. The following section will develop a positive account.

2.2 A Unified Account of Sensible Properties

Here is the dialectical situation. At least most (indeed, as I defend below, all) of our experiences are transparent. The properties they acquaint us with aren't properties of themselves but rather properties of their ostensible objects. But the nature of these properties is mysterious! They are certainly proper properties, not mere modes of presentation. Introspection tells us they are primitive properties but that they don't involve complex rational commitments.

To get at the content of an experience, we have to should analyze the rational commitments it entails. How would it impact our beliefs if we took it to be veridical? A problem in assessing intrinsic phenomenal content is that we are quick to overgeneralize our own case. While it might be intuitive that sensible properties are edenic for instance, we saw it is possible to imagine perceivers that roughly share our phenomenology of color, but who plausibly do not perceive edenic properties, as was the case with the hermit considered above. Differentiating between what is intrinsic to some type of experience and what is true merely about neurotypical human perception or even just average human philosophers is non-trivial. These considerations suggest that a plausible answer must be a minimal representational

[11] Chalmers 2012a, chapter 6.3
[12] Horgan and Tienson 2002.

correlate that not merely all human experiences, but phenomenally similar experiences share generally across possible perceivers. As in the case of the hermit, we can use conceivability intuitions to judge such cases.

So what rational commitments does the experience of *s*-blue entail? With the above considerations in mind, it seems plausible that the intrinsic experiential content of an experience of seeing some object as *s*-blue is *that the object appears a specific way*. Sensible properties are ways of appearing. Call properties of this kind *appearance properties* and the account that claims that sensible properties are appearance properties *the appearance account*.[13]

Note that we can arguably differentiate between *basic* and *complex* appearance properties where the former are primitive ways of appearing, like the way the blue of the sky appears, and the latter are more complex appearances, like appearing to be a cat. While technically complex appearance properties are ways of appearing strictly speaking, for the moment the term appearance properties will refer to basic appearance properties.

Of course, the appearance account does not claim that all we perceive are appearances. When I see a tree in front of me, then it is a tree I see, not a mere appearance. We capture this by holding that perceptual content involves the causal structure of the perceived scenario. We see appearances as caused by the objects of experience. When I see the tree, I see an object that has certain causal characteristics and appears in a certain way in virtue of these causal characteristics. As sensible properties were defined as the properties intrinsically represented by phenomenally individuated experiences this does not entail that causal properties are sensible properties strictly speaking because they are arguably not connected to a specific phenomenal feel.

In this section I will defend the appearance account. First, I want to discuss its relation to a better-known position on the ontology of colour, namely dispositionalism and address some worries that emerge from this discussion. Secondly, I will argue that issues of circularity can be circumvented by adopting a non-reductive notion of appearance. Finally, I will discuss the relation of appearance

[13]The term 'appearance properties' is taken from Egan 2006, who's views however diverge from my own. Alternatively, one may use the term 'secondary properties', which however has the unfortunate connotation that the relevant properties are somehow less real than typical 'primary properties'.

properties, Fregean senses, the notion of self-representationalism, cognitive phenomenology, and metasemantic internalism.

2.2.1 Dispositionalism

There are various similarities between the appearance account and what is called *dispositionalism*. This is roughly the claim that sensible properties are *dispositions* to appear a certain way *under normal or optimal conditions*,[14] while the appearance account claims that sensible properties pertain to appearing a certain way *full stop*. That is, appearance properties carry no direct implications about what would happen normally or under optimal conditions.

In my view, dispositionalism is another example of a phenomenologically thick theory of the content of experience that over-generalizes neurotypical human experience to the intrinsic content of the experience of sensible properties. To see that this, consider a case of *dynamical spectrum inversion*. Eva is subject to a strange and unique condition. She has a non-constant color spectrum. Every thirty seconds, her visual spectrum is inverted. Green things that look green in one moment suddenly turn red and *vice versa*. This might seem horribly confusing at first, but Eva has learned to deal with it just fine. Except in very seldom scenarios her visual acuity is that of a normal adult.[15]

For Eva, the idea that there is *right* or *normal* way objects appear, is alien. It would not make sense for her to decide whether her non-inverted or her inverted spectrum is in fact the "right" one. Therefore, Eva serves as a counter example to the thick phenomenal content predicted by dispositionalism.[16] In other words, believing that a perceptual state is accurate does not preclude that it is not *normal* that things appear this way. Dispositionalism ought to be rejected in favour of the appearance account.

Note that Eva arguably need not be metaphysically possible for this argument to be valid. What we are interested in is content implicit in the bare phenomenology of color.[17] Therefore, roughly being able to imagine *Eva's phenomenology* is

[14] McGinn 1984; McDowell 1985; Smith 1986; Shoemaker 1994a; Kriegel 2002; Burgess 2007.

[15] This does not invalidate the above anti-Fregean argument because it is plausible that Eva experiences changes of properties.

[16] This is true even if Eva were to still live in Eden and Eva believes accurately that it is the edenic properties of the objects that change.

[17] These two ways of conceiving the inverted spectrum argument is pointed out in Thau 2002, chapter 1.

sufficient to realize that there is arguably something wrong about dispositionalism as a theory of sensible colors. We are not contemplating the possibility of a dynamical colour spectrum generally, but merely trying to tease out the rational commitments entailed by things appearing a certain way. And at any rate, it is plausible that one could in principle create a dynamic color spectrum by functionally inverting the right neurological pathways.

I don't want to outright reject dispositionalism. It may well provide a suitable analysis of our public color concepts for example, but it is almost certainly misguided when applied to sensible properties. We are interested in the nature of properties we are presented with in experience just in virtue of phenomenal character, not in the meaning of terms of public language. While dispositionalism may or may not be a powerful lever to connect sensible colors and public colors, the two certainly aren't the same. Obviously, dispositionalism about public colors needs to say more than just that they are dispositions to appear a certain way to normal perceivers because this would make public colors and sensible colors roughly align. In this case, the inverted spectrum argument would threaten dispositionalism about public colors as well. But this flaw might be alleviated by adding that there is some way public colors normally appear to each individual, rather than to a group.

At any rate, it arguably is the confusion of public colors and sensible colors that partly motivates physicalist, dispositionalist and edenic views on sensible properties. The notion, for instance, that colors are there independently of anyone seeing them clearly applies to public colors but, it is far less obvious that it applies to sensible colors. I find it hard to believe that some being could not have been born and raised as a Berkleyian idealist that intuitively holds that *esse est percipii*. Such a being could arguably have normal color phenomenology and therefore perceive sensible colors. But these colors would not seem like they are there independently of one's perception. Similar remarks apply to the idea that it is part of the phenomenology of color vision that sensibly colored things appear the same way to everybody.

In discussing the alleged phenomenological implausibility of dispositionalism, Boghossian and Velleman write:

> When one enters a dark room and switches on the light, the colours of surrounding objects look as if they have been revealed, not as if they

> have been activated. That is, the dispelling of the darkness looks like drawing a curtain form the colours of objects no less than from the objects themselves. If colours looked like dispositions however, then they would seem to *come on* when illuminated.[18]

While not a critique of the appearance account, if plausible, a similar critique could be formulated here. If colors look as though "revealed" then, contrary to the appearance account, sensible colors can't be there merely in virtue of being seen. Further commitments are implied.

I have two comments. First, while there may be a certain plausibility in this observation this plausibility vanishes when we consider that any workable account of the intrinsic nature of sensible properties ought to apply to all agents that share the phenomenology of color vision generally. That is, the *prima facie* plausibility is based on a conflation of intrinsic and extrinsic phenomenal content of color experience. It is certainly possible for there to be an agent that may experience the tomato as *s*-red without involving a commitment how the same object looks independently of this experience. For such an agent, turning on the kitchen light and seeing a tomato looks as though it is now *s*-red, without any commitment on how it looked or was before the light turned on. The cases of Eva or a born Berkleyian may serve as apt examples.

This brings me to my second point. For what it is worth, I am myself such an agent. Having stared at coloured things for quite some time trying to tickle out the phenomenological intuition described by Boghossian and Velleman, I can conclude with confidence (or at least as much confidence as there can be in such matters) that sensible colors really don't appear to me as though they are the way they are independently of whether I look. They just appear the way they appear here and now and are strictly silent on such matters: It seems to me that my color phenomenology is intuitively thin.

Don't get me wrong. Of course, I would be surprised as anyone if, in the blink of an eye the sky would change its sensible color. But this by itself does not show that there are normality conditions built into the very experience of color: They are representations of the causal structure of the perceptual scenario, extrinsic to the nature of color experience itself.

[18] Boghossian 1989, p. 86. Emphasis in the original.

Does this mean that I am a less competent phenomenologist than Boghossian or Velleman or is phenomenology just more variable than these authors suppose? The latter is certainly an interesting psychological question. Be that as it may, the cumulative case composed of my own phenomenological observation and the apparent conceivability of counter-examples, like the born Berkleyian, Eva and the hermit, let me reject the phenomenological claims of Boghossian, Velleman and others.[19]

2.2.2 Are Appearances Reducible?

The claim that sensible properties are ways of appearing is phenomenologically plausible. We will now investigate the nature of appearances in more detail. Among dispositionalists there is a standard distinction between reductive and non-reductive dispositionalism. *Reductive dispositionalism* holds that the notion of appearance can be analyzed as dispositions to cause more fundamental entities, like sensations, sense data or experiential types. *Non-reductive dispositionalism* holds that the notion of appearance is not analyzable in terms of anything more fundamental. Here, that *s*-grey is the property of normally appearing *s*-grey is all that can meaningfully be said about it.[20]

We can make a similar distinction for the appearance account where the *reductive appearance account* holds that appearances are reducible while the *non-reductive appearance account* denies this. The central difference to the correlated dispositionalist views will be that appearance properties are devoid of the implications of assuming phenomenological thickness. Most points made in the literature to refute reductive dispositionalism, except some phenomenological claims, apply to the reductive appearance account as well. As we will see, these arguments decisively refute reductionism about appearances, which is why I embrace and defend a non-reductive variant. We will furthermore see that the non-reductive appearance account is neither trivial nor viciously circular.

I will discuss the reductive appearance account in terms of Sidney Shoemaker's take on sensible colors, but most points I make apply similarly to other variants of reductionism as well. Shoemaker and his followers hold that appearances involve causal relations to experiential types. When I apprehend an object appearing *s*-blue,

[19] See for instance Chalmers 2010; Byrne and Hilbert 2011
[20] The distinction is introduced in Byrne and Hilbert 2011.

I really see the object's disposition to cause a correlated *s*-blue experience in me.[21] We can call the experiential quality correlated with an experience of *s*-blue *s-blue'*.

To see the problems of such an approach we need only ask about the nature of *s*-blue'. Either *s*-blue' is the same property as *s*-blue or it is not. If it is the same property, then we obviously have lost all explanatory surplus of introducing the Shoemakerian account. Appearing *s*-blue is explained in terms of an experience appearing *s*-blue. But why this extra step? When I see an *s*-red tomato, what reason could there possibly be to claim that the tomato appears *s*-red in terms of my experience appearing *s*-red? Why not just claim that it is the tomato that appears *s*-red, full stop? But this would mean that Shoemakers account collapses into a non-reductive variety which we discuss below.

So in order to fulfill its explanatory ambitions the account has to claim that *s*-blue' is some distinct property from *s*-blue and this is arguably how best to interpret Shoemaker. But, upon reflection, it is evident that on *this* reading the account is phenomenologically untenable. When I see the blue sky or a homogeneous blue surface I really am not aware of two properties. Any supposed *s*-blue' that is to express the way my experience feels is precisely the same property that the sky and the homogeneous field bear. The feel that makes the blue experience a blue experience is the very same feel that is attributed to the object of a blue experience. This of course is reminiscent of the Peircean transparency of experience.[22]

I conclude that reductive accounts of this kind are not very attractive. There is no explanatory surplus in accounting for appearances in terms of some kind of mental entity that itself appears in some manner. Rather, appearances are irreducible to anything else. So we end up with the non-reductive appearance account. The simplest, but slightly confusing way to express this view is to say that to be *s*-blue is just to appear *s*-blue.[23]

The most obvious problem with this kind of view is that it may seem viciously circular. One is tempted to ask whether, in order to know whether the sky is *s*-blue, according to the non-reductive proposal, I did not first have to figure out whether the sky *appears* *s*-blue. But then our definition tells us that, in order to assess

[21] Examples are Shoemaker 1994a, Shoemaker 2003, Kriegel 2002 and Kriegel 2008.
[22] This is a variant of the argument found in Boghossian 1989.
[23] Non-reductive versions of dispositionalism are defended in McGinn 1984, McDowell 1985, Smith 1986 and affirmatively discussed in Burgess 2007.

this, one needs to figure out whether the sky appears as though it appears *s*-blue because being *s*-blue just means appearing *s*-blue. And thus, one may conclude, the non-reductive proposal would keep us from ever seeing any sensible colors at all.[24]

Issues of circularity are highly non-trivial. J.A. Burgess comments:

> It must have come as great surprise and disappointment to the philosophical zealots who sought out circularity in the works of their colleagues with the enthusiasm and compassion of bounty hunters when, in the late 1970s, philosophical logicians began talking as though circularity might not always be a defect. By the mid-1980s a few had even been brazen enough to say as much in print. But although the thought that circularity could be benign became commonplace, attempts to say when it is benign and when malign were conspicuously, sometimes spectacularly unsuccessful.[25]

So how shall we evaluate the claim that the non-reductive analysis is *viciously* circular? First, to see that circular analysis isn't, by itself, a *logical* issue, consider "*s*-red is *s*-red". While this is of course not informative, it certainly isn't a logical mistake either. That is because it is true. Circularity all by itself isn't something to worry about.

How can a circular analysis be informative? To take an example used by Boghossian and Velleman, "Courage is the disposition to behave courageously" is arguably logically non-contradictory in the same sense as "Courage is courage" is. And it is obviously not empty because it tells us that courage is a certain behavioral disposition. Therefore, circular analysis can be informative, too.[26]

So what *are* the reasons to be sceptical of circular analysis? Circular analysis is certainly *pedagogically* useless. Knowing that courage is the disposition to behave courageously isn't helpful if one does not already have the concept of courage. Therefore, non-circular analysis is to be preferred to circular analysis whenever possible. If a concept is primitive however and cannot be defined in terms of anything else, circular analysis might already be the best thing we can come up with. If courage were a primitive way to behave that cannot be analyzed, "Courage is the disposition

[24] Similar doubts about non-reductive dispositionalism are to be found in Sellars 1956, Boghossian 1989 and McGinn 1996.
[25] Burgess 2007, p. 216
[26] Boghossian 1989.

to behave courageously" may be a perfectly fine result of an investigation into the nature of courage. Sometimes a circular analysis may be the best thing available.

Note that, if this is correct, the analysis of appearances is precisely where one would expect to run into circularity. Appearance concepts famously cannot be communicated to anyone who does not already possess them. If one does not know the property of *s*-red, no amount of talking will help that person to grasp it. Therefore, the circularity of any suitable analysis of *s*-red is to be expected.

Is the non-reductive analysis of appearances informative, in spite of its circularity? Certainly. It tells us that sensible properties are primitive ways of appearing. So while they can't, by their vary nature, be exhaustively defined, this still is a substantial insight into the nature of sensible properties. Readers that are still not fully convinced should bear with me until chapter three, where the notion of appearance properties will be employed as a central puzzle piece in the explanation of conscious experience in physical terms. There our inferentialist analysis will throw light on the question of why exactly sensible properties cannot be defined in terms of anything else.

But of course not all circularity is unproblematic. According to a helpful distinction by I. L. Humberstone we have to differentiate cases of mere *analytical* circularity, where the same concept shows up on the left and the right-hand side of an analysis, from *inferential* circularity. Inferential circularity ensues if the application of the right-hand side of an analysis presupposes the application of the left-hand side.[27] This leads to a regress that makes it impossible to apply the analysis at all.

Above we worried whether the non-reductive analysis is inferentially circular in the sense that, to know whether something is *s*-blue we have to know whether it appears *s*-blue. But to know whether it appears *s*-blue we have to know (by substitution) whether it appears to appear *s*-blue, *ad infinitum*. Thus, to start seeing that the sky is blue I have to engage in an infinite loop of concept applications, which I cannot do. So if the non-reductive appearance account is correct, no-one could ever see the color of the sky, or so it seems.

To see the flaw in this reasoning, we have to differentiate between a transparent (not to be confused with phenomenological transparency discussed in the previous chapter) and an opaque understanding of the predicate "to appear...". *Transparent*

[27] Humberstone 1997; Burgess 2007.

contexts are such that substitution of terms with equal referent will preserve truth-value. An example would be the sentence "John Dee was Queen Elisabeth's court astrologer." If we substitute "John Dee" by "The first man to be called 007" (which is John Dee), the truth-value will be preserved.

However, if we compare this to the sentence "Susanne believes that John Dee was Queen Elisabeth's court astrologer.", substitution will not preserve truth value, because this depends on whether Susanne knows that John Dee was the first man to be called 007. Thus, "Susanne believes..." creates an opaque context, such that substations of referentially identical terms are not guaranteed to preserve truth-value.

In the argument for the circularity of the non-reductive appearance account, a transparent reading of "to appear ..." is assumed, because otherwise the substitution procedure will not be valid. It would not be the case that to appear s-blue can be substituted at will with appearing to appear s-blue. But if the transparent reading is assumed, then the result of an infinite nesting of appearances is actually not threatening at all. This is because, on this reading, to appear to appear... s-blue *just is* to appear s-blue. So what one needs to do to apply the infinite set of concepts is just the same as applying it once. To see that the sky appear to appear... s-blue one just has to see it as s-blue.[28]

Let's sum up our discussion of circularity. It is, first of all, a myth that circularity is by its very nature logically bogus. It may be pedagogically useless but this on its own should not be held against a philosophical thesis. To show that the non-reductive appearance account is viciously circular, an opponent needs to show that it implies that an inferential loop arises whenever an appearance concept is applied, i.e. that it is inferentially circular. But this demonstration will run through if one tacitly conflates two interpretations of appearance-talk. When we decide for one reading the inferential loop either does not ensue or it is harmless: If we interpret appearance opaquely, the substitution needed to get the loop off the ground is illegitimate. If we interpret appearance transparently, the loop is non-threatening because every iteration of the loop puts the same demand to the perceiver. I conclude that while the non-reductive analysis is circular, it is not viciously so.

[28] Byrne and Hilbert 2011.

A final concern that may arise is that a non-reductive analysis is incompatible with naturalism about perceptual states. For appearance properties seem to be primitive entities that now cannot be explained in terms of anything naturalistically acceptable. This charge isn't easy to refute and it of course relates to the question whether consciousness is naturalistically explainable. Here I will have to refer the reader to part two and three of the thesis. It will turn out that a naturalistic understanding of such properties is possible, but only if we make some concessions regarding what it means to be a naturalist.

2.2.3 Some Features of the Appearance Account

Let's finish our discussion by briefly addressing some features of the appearance account of sensible properties.

Modes of Presentation

Appearance properties resemble modes of presentation in many ways. We saw that one natural representationalist response to the challenge of the inverted spectrum is to claim that Invert actually represents the same physical color properties as Nonvert, but that the associated modes of presentation have been switched. This however, barely squares with the phenomenology of color vision because the phenomenology of seeing s-red and seeing s-green is a difference in the *properties* of their bearer.

The author that seems so far most sensitive to the dual role of sensible properties as properties and, at the same time, similar to modes of presentation, is David Chalmers. The upshot of his edenic view is a form of content pluralism where conscious states bear both a Russelian content that involves edenic properties, and a Fregean content, that involves the properties in the actual world that most closely parallel the role of the relevant edenic properties in the actual world. As Chalmers motivates the Russelian component of content with reference to thick phenomenology this kind of pluralism is ultimately implausible.

Standardly, modes of presentation are thought of as more basic entities than properties. However, we can also conceive of properties that closely parallel the function of modes of presentation, like the property of appearing first at the evening sky and the property of disappearing last from the morning sky. Representing Venus

as the morning star means representing it in terms of the former property, representing it as the evening star means representing it in terms of the latter. My claim is of course not that modes of presentation are properties, just that there is a way to single out properties such that there are properties corresponding to many modes of presentation. As properties themselves can be represented under modes of presentation, the claim that all modes of presentation are just more properties should be handled with caution.

It is natural to suggest that appearance properties are precisely a kind of property that functions like modes of presentation of external objects in perceptual experience. By representing objects in terms of their appearance properties, an agent represents these objects partly in terms of their relation to her own perceptual system. Consciously seeing the tomato as *s*-red does not merely mean representing something about the nature of the tomato, but it also tells us something about the representational state that represents the tomato.

Note that in a similar way one may argue that appearance properties really aren't intrinsic to the object of experience but are relational properties of object and perceiver. However, the distinction between relational and intrinsic properties is far less clear than I would like and there arguably can be no neat dividing line be drawn here.[29] Still, in the coming discussion I will at some points rely on the relationality of appearance properties. It is important to keep in mind that appearance properties are not relational in the sense of being reducible to a causal relation to an experiential type, as the reductionist would have it, but in the sense of always involving an intentional object *and* a representational system. In this sense appearance is a primitive relation.

The Reflexive Character of Consciousness

This brings us to the second feature of the appearance account, namely that it involves a tacit commitment to a form of self-representationalism. This view is sometimes put as the view that all phenomenal consciousness involves a degree of self-consciousness. More exactly, *self-representationalism* is the view that phenomenal properties are *reflexive representational properties* where representational properties are reflexive if the associated representational states are among their own intentional objects.[30]

[29] For a very brief overview of some difficulties, see the discussion in Seager 2006.

[30] A popular defence of this view is to be found in Kriegel 2009, a valuable overview is given in McClelland forthcoming.

Self-representationalism can be motivated phenomenologically in that consciousness can typically be separated into two different aspects. First, there is a *qualitative aspect* in that consciousness involves awareness of the qualities of some typically non-mental object. When one sees a blue rose, one is conscious of the properties of the rose. On the other hand, consciousness involves a *reflexive aspect* in that consciousness involves the fact that the object appears *to the subject* (what is sometimes called the *subjective character* of consciousness). When one sees a blue rose, the rose appears to the subject and this fact is itself part of the experience's phenomenal character. Self-representationalists typically hold that the qualitative aspect of experience is to be explained in terms of the non-reflexive representational properties of a conscious state, while the reflexive aspect is to be explained in terms of the reflexive representational properties of that state.[31]

An interesting corollary of the appearance account is that it implies a form of self-representationalism. If conscious perceptual states represent how things appear, and appearance is a representational notion, then conscious perceptual states necessarily represent an aspect of themselves, namely *how* they represent their object. Appearance is a partly representational notion. If sensible properties are primitive ways of appearing (i.e. of being represented), then phenomenal content refers back to the very representational state itself. Note that the point here is not that sensible properties themselves are reflexive, but representational states representing them are. These states, over and above attributing properties to a perceptual object, will attribute certain properties to themselves.

These issues are too complicated than would we wise to elaborate here. We will return to these issues in chapter eight, where we will argue for the phenomenological plausibility of a reflexive aspect of consciousness and also see that the qualitative and the reflexive aspect can sometimes come apart. For now, it is sufficient to see that the appearance account entails a form of self-representationalism.

The self-representational aspect of perceptual phenomenology is hard to account for in terms of a naturalist account of representation. Giving such an account will be a central issue of chapter eight where we will discuss the inferential architecture underlying transparent self-knowledge.

[31] Kriegel 2009, p. 45-57.

Cognitive Phenomenology

That sensible properties are appearance properties may also apply to cognitive phenomenology. Above, I mentioned that the problem of edenic properties also threatens to make cognitive phenomenology largely falsidical. If I associate the function of addition, say, with some primitive edenic feel, then this would just be a mistake. Addition is an abstract object that certainly doesn't have such a feel as one of its characteristics. But if the feel of addition is an appearance property, no such worries arise. While the function of addition might not be associated with some primitive edenic feel, it is most probably associated with a appearance property that denotes how addition appears to a given subject.

This of course is not meant to be a theory of cognitive phenomenology, but a bare proof of principle that the appearance account is not in conflict with such phenomenology and its supposed explanatory value. Defenders of cognitive phenomenology and its explanatory value in content-determination should prefer the appearance account over the edenic account.

Internalism

Plausibly, the view that sensible properties are appearance properties correctly localizes the supervenience base of the phenomenally conscious content. If content were somehow determined in the way tracking representationalism supposes it is, namely in virtue of the statistical correlation between the representing state and the represented state, then representational content could not be determined locally by what happens inside the skull. To take Ned Blocks famous example, a physical duplicate of me, growing up on an *inverted earth*, a planet where all surfaces have the opposite color from the one on earth but who has color inverting lenses implanted into his eyes, would arguably see the world just like I do. But the physical colors tracked by his experiences would be different. Thus, according to tracking representationalism, they should be phenomenally different experiences. Tracking representationalism falsely locates the supervenience base of experience.[32]

If, as I contend, color experiences do not represent physical properties then nothing stands in the way of supposing that the relevant representational content

[32]Block 1990. Arguments to the same end can be found in McGinn 1997 and Pautz 2006.

is determined entirely by what happens inside the skull. As it seems *prima facie* highly plausible that phenomenal properties are fixed by what is internal to the skull, this is a win for the secondary position over the tracking view. What is lacking of course is an explicit metasemantic account that explains how appearance properties come about. This will be subject of part two.

2.3 Summary

According to the appearance account the properties we are directly aware of are appearance properties. In this section we have primarily dealt with the nature of these appearance properties. We have defended that they aren't phenomenologically implausible, because intrinsic phenomenology is thin. We have seen that the circularity involved in appearance properties isn't vicious, but a benign consequence of their irreducibility. In the coming chapter, we will complete our argument for representationalism by discussing various counterexamples.

3 Defending Representationalism

In this section we will discuss putative counterexamples to representationalism. In introducing the transparency argument I already defended that there is a general strategy for arguing for representationalism. If a certain set of experiences elicits Peircean transparency in the sense that the properties they directly present us with are properties of the object of experience, then representationalism is arguably true for that set of experiences. What remains is a proof that indeed all experiences are transparent, not just perceptual states, that is, a proof that all properties we are aware of in experience generally are properties of the objects of experience.

There exist a variety of different counterexamples to representationalism that, in the present context, we can largely frame as counterexamples to Peircean transparency. I will now go through three broad categories, namely affective experiences, the effect of attention on perception and cognition and finally the unity of consciousness. I will show for each case why it does not constitute a counterexample to transparency of the form the anti-representationalist requires. While they do not refute representationalism, we will see that there are important lessons to be drawn from these cases.

Representationalists may be tempted to refute putative counterexamples to Peircean transparency on ground of different manners of representation. For instance, there is a difference in visual phenomenology when I am paying attention to the object of experience as opposed to when I am distracted. Now one may hold that this phenomenal difference, while it is not a difference in content strictly speaking, still corresponds to a difference in manner of representation. One may hold that the attended experience represents the object of experience *saliently*, while the non-attentive experience does not represent it this way.

Remember that representationalists that try to account for phenomenal properties purely in terms or representational content are called pure, while impure representa-

tionalists appeal to manners of representation. Thus putative counterexamples to representationalism may motivate impure representationalism. I will try to avoid this manoeuvre and defend pure representationalism wherever possible. The reason for this will become fully transparent only in later chapters where I will appeal to theories of cognition that account for attention, affect and unity of consciousness in pure representational terms, i.e. in terms of content alone. Thus the current discussion may thus serve as a defence of the phenomenological plausibility of such theories.

3.1 Affective Experience

Probably the most popular putative counterexample to representationalism is pain. There are at least three different ways in which pain might fail to be transparent. First, it might be denied that it has any content or intentional object at all. This may be supported by the fact that we normally do not associate pains with conditions of satisfaction. There is no such thing as a true or false pain, at least in the sense required: A pain without a clear bodily cause is just as 'true' as any other.

To deny that there is a content to pain experiences at all is too radical. That pains *do* have an object, as representationalists like to point out, is evident in the fact that it normally is associated with a felt location. If you bump your knee, the associated pain is not an undifferentiated blob but a phenomenologically richly textured experience. The pain may be throbbing or stabbing, dull or bright, more or less intense and has a more or less definite location in your body image. The most natural explanation of these facts is to hold that pain experiences have an intentional object of some kind.[1]

A second kind of worry is that it is unclear at best what kinds of properties experiences of pain present us with. But such worries are again fueled by the tracking picture of representation. What do pain experiences track? Tye has suggested that they track and represent bodily damage.[2] However, it seems phenomenologically implausible that this kind of content is what we are conscious of when we are in pain. Arguably you can have the bare experience of pain even if you don't even know you have a body. Pain does not present us with bodily damage. It presents us with raw *painness*.

[1] Shoemaker 1994a, Tye 1995.
[2] ibid., p. 113.

This kind of worry evidently is a variant of the problem of the phenomenological implausibility of the tracking view, and they can be dealt with in a similar manner as above. The properties that we are experiencing when we are in pain are primitive properties. In particular, they are appearance properties that pertain to the way the object of the pain experience appears. This absolves us from giving some obviously phenomenologically inadequate analysis of pain experience. The pain represents its object exactly as it appears to.

Thirdly and finally, one may deny that the object of a pain experience is distinct from the experience itself. When I hurt my knee, then it is the pain that I experience, not the knee. This may be supported by the observation that having an experience of pain is all there needs to exist for the pain to exist. Even if someone does not even have a knee, we normally would not say that this implies that one could not possibly have pain in the knee (as in the case of phantom limb pain).

It is unclear at best whether this third and final contention really establishes that that pain experiences aren't transparent. To show this, I will argue that the intuition that pain does not have a distinct intentional object is grounded primarily in the way we speak about pain, rather than in its actual phenomenology and phenomenal content.

The contention that the object a pain experience is the experience itself needs some explication and motivation. To get some grip on these characteristics of pain we may contrast its grammar[3] with the grammar of appearance talk. By speaking of appearances, we may pick out the *ostensible* object of an experience as opposed to its actual object. If I hallucinate a bird flying by, then I can talk about this in two different ways. Either I (falsidically) say that there is a bird flying by or I (veridically) say that it appears to me as though there was a bird flying by.

There is an important parallel here between pain-talk and appearance-talk. Just as in the case of pain in the knee, appearance talk may even be veridical if *there is no* object of experience, as in the case of the hallucinated bird. Both pick out the *ostensible* object of an experiential representational state. This becomes especially evident when we contemplate that an *appearance of pain* and a pain seemingly are just one and the same thing. If something appears to be painful it arguably just *is* painful.

[3]I use the term "grammar" in a Wittgensteinian sense of what are the meaning-conferring norms that govern the usage of the term.

This suggests a radical claim. The difference in an experience of pain and the experience of something as green is not primarily a deep difference in phenomenological structure. Both attribute certain appearance properties to their ostensible object. The difference amounts to the way we individuate experiences in terms of our language games. Grammatically speaking, saying that I am in pain is rather like saying that it appears to me as though something was a bird, rather than like saying that it *is* a bird.

We may imagine a language game where pain talk can not be analogized to appearance talk in this way. In such a game, an experience of phantom limb pain is just as illusory as the hallucination of a bird flying by. The important point is that speakers playing this hypothetical language game need not have experiences that are any different from the experiences we have. When they bump their knee they may exclaim: "Ouch! It appears so painful!" If this is indeed a conceivable scenario this illustrates that the difference between pain and perception, regarding their intentional object, is not a danger to representationalism. It is merely grounded in linguistic convention and therefore extrinsic to the representational structure of consciousness.

It is natural to ask *why* pain is individuated differently from perceptual and cognitive states in our language games. The reason arguably has to do with its biological function. Pain is a highly *affective* state that indicates bodily damage to the organism.[4] To do this, the experience of pain has to be salient and hard to ignore.

Language cuts along joints of relevance for its users. Because pain is highly affective, an experience of pain has relevance even if its ostensible object does not exist. If someone has horrible pain in her phantom limb that makes her incapable of work, this is almost as relevant for social practice as pain that is caused by real injury. Thus, if there were no word for pain that abstracts away from the veridicality conditions of experience we probably ought to invent one. If my view that differences between pain and perception are grounded in our language rather than the structure of the experiences themselves, then pain arguably does not offer a counterexample to representationalism.

Just as pain, joy can serve as a counterexample to representationalism. All three points mentioned above apply. At a first glance, one may suspect that our joy experiences do not have an object at all. If they do have an object, it is not quite clear what

[4]Note that this does not imply, as the tracking view holds, that this is the phenomenologically manifest representational content of pain.

kind of properties are presented in joyous experiences. Finally, one may be tempted to suggest that joy does not refer to an independently existing intentional object.

We can use our treatment of pain as a template for dealing with joy. It is phenomenologically implausible to claim that joy does not involve representational content at all because joy is phenomenologically rich and can normally be localized in specific parts of the body. Furthermore, joy is normally directly associated with other contents of experience. One does not experience some sensation *and* joy. One experiences sensations, impressions and the things of this world *as* joyous. Thus, it is incorrect to say that joy is non-representational.

What does joy represent? Apart from causal structure (apart from felt location there normally is a particular object that is experiences as the source of joy), felt location and so on, joy arguably represents appearance properties.

Finally, just as pain, joy is highly affective which is why we have a way of individuating it independently of the commitments to its "veridicality". This is why, when a hallucinated experience is pleasurable, it does not make sense to say that the pleasure is non-veridical: For pragmatic reasons this turned out to be the way our language functions.

The argumentative strategy in this section was inductive. Rather than giving an argument from first principles that shows that all experiences need to be transparent and representational, I argued that pain and joy, which may seem like counterexamples at first and maybe second glance, can be reasonably argued to be purely representational phenomena.

The argumentative strategy may be used as a template to refute other counterexamples similar to affective experience. Experiences where we normally abstain from drawing distinctions between appearance and reality can be properly rendered purely representational by considering their rich phenomenology and felt location. That their content is elusive can be captured by an appeal to appearance properties. And finally, the fact that we do not draw an experience-object distinction can be explained as extrinsic to the actual phenomenology.

3.2 Ways of Attending

There is another family of putative counterexamples to transparency and representationalism that may broadly be classified as issues regarding attention. Experiences may differ phenomenologically depending on the degree of attention we pay to them. This difference, the enemy of representationalism may insist, cannot be construed as a difference in the object of experiences. Therefore, the difference must be a difference in non-representational properties.

A particularly interesting way of illustrating the anti-representationalist argument is the distinction between the periphery and the center of attention. There is no question of whether there is a phenomenal difference of seeing a bird flying by and paying attention as opposed to seeing a bird flying by peripherally. However, the argument goes, this is not a difference *of the bird*! Thus, it seems cogent to conclude that it must either be a difference in the manner of representation or a difference in the mental paint involved. At any rate, as my shift of attention does not change the bird, it seems that any other construal of the situation would involve the thesis that shifts of attention involve a perceptual illusion.

A possible strategy for dealing with these cases may be to move to impure representationalism by explaining phenomenal differences involved in shifts of attention as differences in manner of representation. But our set task is to answer whether counterexamples can be refuted on pure representationalist grounds, i.e. by appeals to content. We will later see that plausible cognitive models of perception conceive attention as a second-order representation of first-order perceptual representation that represents first-order model quality. Here the central difference between center and periphery of attention is that representations at the center of attention are represented to be more reliable than peripheral representations by a second-order model. Such accounts evidently deal with attention in a purely representational manner i.e. differences in attention are rendered as differences in representational content. Here I want to focus on the question whether such a purely representational account is phenomenologically plausible.

If the construal of attention as a result of second-order modelling is to be phenomenologically plausible then we have to ask ourselves whether we can conceive of an attentional difference as a difference in how the object of experience is represented to

be represented. Now it seems to me that such an account indeed is intuitively plausible. To see this, note that, just as it would be strange to suppose that the phenomenal difference of seeing a bird focally versus seeing it peripherally as a difference of the bird's intrinsic properties, it is also hard to deny that something about the bird's relational properties changes, relational that is with regards to a representational system. The second-order representation model of attention arguably can address both aspects: It explains why shifts of attention do not effect first-order properties of the intentional object while effecting its second-order properties. Thus a representationalist account of attention cannot be easily refuted on phenomenological grounds.

One line of argument that employs differences in attention to argue against representationalism is developed by Block. Perceptual science shows that attention can lead us to misjudge certain visual features. Focus your eyes on the black dot in the middle of the graphic below. Then let you attention wander to the right, while leaving your gaze steady. After a moment, you may realize that the grating on the left is misjudged as being as pronounced as the grating on the right. However, in reality, both gratings differ. This is a simple example of an attention-induced misrepresentation and by itself not a threat to representationalism: It is just an ordinary illusion.[5]

The same effect would arguably occur if there *was no* grating on the right side of the fixation dot. That is, your attention would also change how pronounced the grating appears if there were no other grating you could compare it to. But this, Block says, is problematic for the representationalist because there is no standard for settling the question, which of the two is *right*, the more or the less pronounced perception. But then there are no conditions of satisfaction associated with the change of pronouncedness and thus no content. It seems we have found a change of phenomenal properties without a change in content.[6]

Blocks reasoning is ingenious but fallacious. It is true that it would be strange to decide which of the two perceptual situation is veridical. But Blocks argument misses the fact that the anti-representationalist understanding is equally worrisome, namely that the difference in experience does not pertain to the object of experience! Evidently it is something *about the grating* that changes when we attend to it. But

[5]The image is taken from Block 2010. Thanks to Ned Block for the permission to use it.
[6]Ibid.

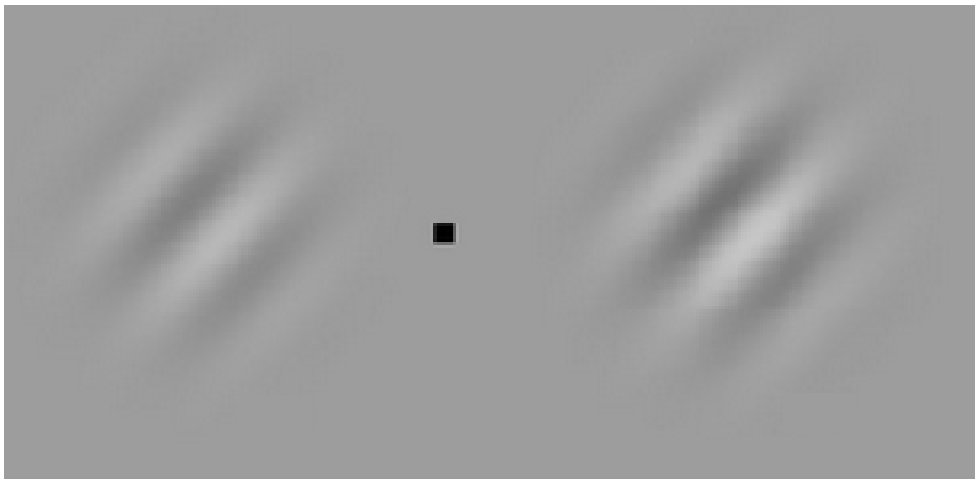

Figure 3.1: An illustration of the warping effect of attention on perceptual contents. Focus on the black spot. Now shift your attention to the left. After a little practice you will find that the attention affects the pronouncedness of the grating such that both, left and right, now appear equal. Crucially the effect will arguably also ensue if there is no second grating for comparison and then there is no intuitive way of telling which appearance is 'correct'. Block takes this to be an argument for the view that consciousness need not have conditions of satisfaction.

facts about the intentional object of a state are just facts about its content. That is to say, the experience we are dealing with seems transparent in a Peircean sense.

Luckily the two intuitions are reconcilable in terms of a theory that conceives of attention as a second-order model. On the one hand we can explain the intuition that shifts in attention do not effect the intentional object: They do not effect the object of experience. On the other hand we can explain why there is a difference in how the object appears: There is a difference in how the object is represented to be represented, i.e. in its relational properties. There is no deep problem involved in accounting for attention without invoking non-representational mental paint.

3.3 Representationalism and Unity

There is an intuitive sense in which each of our conscious experiences is one. It forms a unity. When I see the rain pattering against the window and taste fresh coffee at the same time, then these typically aren't two streams of experience. Instead, they are somehow integrated into one and the same experience. They are *phenomenally unified*. Intuitively, we possess a *phenomenal field*[7] of experience.

[7]Coined in Bayne and Chalmers 2003.

From the standpoint of pure representationalism, the existence of the phenomenal field may seem puzzling. A pure representationalist has to explain unity in terms of the content of experience. However, the point of the phenomenal unity of the taste of coffee and the sight of the rain pattering against the window is precisely that these objects are not one. They are not essentially unified. Thus, there seems to be a phenomenal fact here, namely that experiencing coffee while experiencing rain and experiencing rain while experiencing coffee is different from experiencing either one on their own, that can not straightforwardly be explained in terms of any pure representational fact. One may therefore suspect that the right account of phenomenal unity will explain it in terms of manners of representation or mental paint, rather than in terms of content.

In this section we will see why this is not so and why pure representationalism can accommodate unity. My strategy will be similar as before. First, I will argue that the supposedly non-representational feature of experience is really representational, after all. Then I will argue that we can explain it in terms of content alone rather than manners of representation.

Let us start by spelling out the notion of the unity of consciousness in more detail.[8] I take it that the intuition that consciousness is somehow unified is quite salient and clear. When I look out of my window I see bees hovering over tomato blossoms, the cloudy sky above and hear the nearby street noise, these can be unified into one. These experiences don't merely happen after or, whatever this would mean, besides one another. They are parts of one and the same experience, and they constitute one phenomenal character. This is why Tim Bayne and David Chalmers call this phenomenon the phenomenal field.[9]

What is the nature of this phenomenal field? The first step in understanding phenomenal unity is to differentiate the relevant sense of unity from two competing notions. First, as I already said, phenomenal unity ought to be differentiated from the superficial unity of the intentional object. The brown color of my coffee mug and its shape aren't dissociated fragments, but they appear as properties of my

[8]Here, I will rely on the work of Bayne and Chalmers. (ibid.)

[9]We will restrict our discussion to *synchronic* as opposed to *diachronic* unity, i.e. we will discuss what unifies experiences occurring at one and the same time into one, rather than asking what connects experiences occurring after one another within the same stream of consciousness. I take it that diachronic unity, while being interesting phenomenologically, does not involve any additional challenges to the representationalist's point of view.

cup. The cup experience is *objectually* unified. However, I may experience a pain in my knee while seeing my coffee mug. These are two objects that, on the surface, aren't represented as unified into one super-object. However, the experience of them *is* phenomenally unified in that both are part of one phenomenal field. Thus, phenomenal unity can be meaningfully distinguished from objectual unity.

Phenomenal unity cannot easily be cashed out in terms of the unity of the intentional object. What positive account can we give? Bayne and Chalmers argue that a reasonable way of explicating unity is in terms of *subsuming* states. On their account two phenomenally conscious states are phenomenally unified if there is a phenomenally conscious state that subsumes them both. In other words, if phenomenally conscious state X and phenomenally conscious state Y are phenomenally unified, then there is a further state Z that may be described as the way it is like having X and Y together.

The notion of *subsumption* here is understood as a primitive that expresses the fact that if Z subsumes X and Y that Z has "phenomenal parts" X and Y. Note that this subsumption relation does important conceptual work. If we erase the notion of subsumption and try to define phenomenal unity of X and Y in terms of some state Z that captures what it is like having X and Y together, this would be regressive.[10] For we can now ask the question of what it is that unifies X, Y and Z. The subsumption approach would answer that this is achieved by the subsumption relation holds between these phenomenal parts.

As I have already hinted at, if we conceive of phenomenal unity subsumptively, this results in a problem for pure representationalism. The only kind of unity a pure representationalist may accept is some kind of unity of what the state is about, i.e. unity that is wholly determined by pure representational properties. The pain I experience while seeing my cup is not experienced as part of the cup, but still the experience of pain seems to modify the experience of the cup as the cup experience gets subsumed into a larger whole. While the phenomenology of the experiences by themselves may be a matter of content, the fact that the experiences

[10]See Tye 2003. For an attempt to reject the regress claim, see Bayne 2010, p. 30-32. For an attack on Bayne, see Wiese 2018, p. 43-44. I will not enter the discussion in detail, because my own solution, to be elaborated below, explains the phenomenological datum of unity in terms of content and therefore doesn't rely on accepting or rejecting any mereological claims. That is, if we can account for the explanandum in terms of content, then no explanation in terms of mereology of mental states is required.

are part of a subsuming experience cannot be. In the remainder of this section, I want to demonstrate why this reasoning is unconvincing.

The first step in seeing that the reasoning against pure representationalism is fallacious is to see that the subsumptive approach to unity is not without alternative. As Tye has pointed out, just because one unified representational state can be described as having a variety of contents, this does not imply that the representational state is composed of smaller representational states. This is seen easily by contemplating an example.

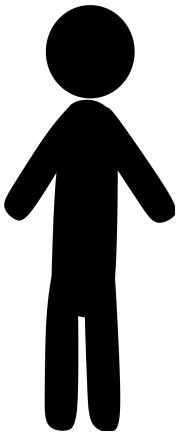

Figure 3.2: Illustration of the holistic properties of many representational states. The image of a man is not composed of images of arms and images of legs but of mere strokes of paint. The question of how the representational parts form an representational whole does not arise. Something similar, I suggest, is true in the case of conscious representational states.

The above figure obviously shows a man.[11] It also shows the man's head, and it shows the man's torso. However, we are not dealing with an representational state that is composed of smaller representational states! In particular, the question of how representations of the man's torso and the representation of the man's head combine into the representation of a person is confused. We are not dealing with a complex representation built from simple representational parts, but a complex representation built from *non-representational* parts.

Phenomenally unified representational states aren't constructed from simpler representational states. Their content is holistic in that there is only a whole representational state with a single representational content.[12] Note that this claim involves a substantial commitment regarding the nature of mental states, a commitment that

[11]This image is available under a creative commons licence on commons.wikimedia.org.
[12]This position is developed in Tye 2003, chapter 1.

will play a role in the theory construction of the following chapter: We have to assume that representational states are holistic instead of atomistic in the way required.

On the resulting view, phenomenal unity and pure representationalism are not mutually contradictory. We do not have to conceive of unified consciousness as a sum of it's parts, and therefore we do not have to explain how representational parts form a complex representational whole. However, this still does not give us a positive account of the phenomenological datum of phenomenal unity. How is the phenomenological fact that the objects appear to us in a unified manner to be explained?

I think the most plausible account of the phenomenological datum of unity can be given in terms of what is experienced, i.e. in term of the content of experience. To see this, first let us again consider a standard example of objectual unity. I see my brown cup in front of me. The experience is objectually unified: the brownness and the shape aren't floating around as dissociated fragments. They are experienced as aspects of one and the same thing. But also, the experience is phenomenally unified: Experiencing the brown cup is different from experiencing brownness and shape as fragments.

Now the crucial question: Are there two kinds of unity here? I don't think so. The phenomenal unity of the experience of shape and the experience of color *just is* the unity of the cup, as experienced. I wouldn't know how to conceive of the experience of the cup as phenomenally unified in the way it is, without the experienced cup to be objectually unified as well. Similarly, the idea of an experience of a cup that is objectually unified but not phenomenally unified seems meaningless.

There is an important lesson to be drawn here. Some instances of phenomenal unity can be rendered as varieties of objectual unity. As a result one may ask whether, contrary to our first impression, phenomenal unity generally can be explained as a form of objectual unity. But as I already argued this seems *prima facie* implausible in many cases. When I see the cup while I am in pain, both experiences are phenomenally unified. However, the cup and the pain aren't represented as part of one super-object.

While it is true that, in some sense, cup and pain don't form a super-object, there is also a sense in which they do. Of course I experience cup and pain *in relation to each other*. Objects of experience are typically experienced as inhabiting certain contexts and different objects can occur in one and the same context. For instance, I experience the pain at a definite location in space and I experience

the cup at definite location in space. I experience them at a particular position in time, too. And finally, I experience them from a singular perspective, i.e. in relation to a singular observer. All these contextual embeddings are aspects of the rich content of neurotypical sensory experience.

Thus, my solution to the problem of unity is this: The only kind of phenomenal unity there is, is the unity of what is experienced. While, what we might call *superficial* objectual unity merely unifies singular objects into one, there is a *background unity* that is constituted by the fact that all elements of experience are experienced in contexts that can also mutually include each other, and thereby the objects of experience are experienced in relation to each other.[13]

This solution makes it plausible that the popular metaphor of the field-nature of experiential unity ought not to be understood as being grounded in consciousness itself having some kind of field-like structure, whatever this might mean exactly.[14] Rather, I take it that the field-like nature of experience is merely the correlate of our intrinsic model of space. Our minds naturally order the contents of experience in a spatial matrix and thus we are disposed to think of consciousness as field-like.

Chalmers and Bayne deny that phenomenal unity could be explained in terms of the experienced spatial relations of the objects of experience. Moods and feelings, they claim, are not experienced in space at all.[15] I take this claim to be phenomenologically dubious. Normally, most moods and feeling are associated with distinct parts of the body, though this may not be intuitively obvious if one has never investigated their phenomenology in any detail. At any rate, even if it should turn out that it is correct that there are non-spatial concurrent experiences, these probably would still be experienced as happening in some temporal relation to each other. This of course relates to the second popular metaphor of unity, namely the Jamesian metaphor of the *stream of consciousness*.[16] To show that background objectual unity cannot explain phenomenal unity one would have to show that there are cases of phenomenal unity without any kind of spatial or temporal unity. It seems that such a scenario is indeed inconceivable.

[13]This coheres nicely with the critique of conceptions of phenomenal unity that define it mereological terms given in Wiese 2018, chapter 3.

[14]Searle's musings on 'unified field model' of consciousness (Searle 2000) may be based on such a confusing of the content of consciousness with its realizing substrate.

[15]Bayne and Chalmers 2003.

[16]James 1950, p. 239.

My discussion so far has been neutral on the question of how wide-spread phenomenal unity actually is. All I have assumed is that it sometimes occurs, and asked the question whether this is a problem for pure representationalism. Different philosophers have taken different views on the actual extend of phenomenal unity. Thomas Nagel has argued, that, while phenomenal unity is the default for conscious experiences in the neurotypical subjects, split brain cases offer some support for the thesis, that our concepts of phenomenal unity can be inapplicable in some cases.[17] Bayne and Chalmers have argued that split brain cases can be interpreted as breakdowns of access unity without a breakdown of phenomenal unity and voice the suspicion that consciousness may be necessarily unified.[18] Tye denies this, claiming that local breakdowns of unity may happen in special circumstances.[19] Still others, like Daniel Dennett, have argued that we are deluded in our sense that consciousness is normally unified.[20] In later chapters, we will tentatively side with the latter authors in that we will assume that the wide-spread impression of phenomenal unity is in fact a result of an introspective limitation. We will give a speculative account of how the unification of representational content typically results from the action-oriented nature of conscious mental functioning. The account will give us some reason to suspect that our very attempt to introspectively observe the contents of our mind tends to collapse it into a unified whole that did not always exist prior to the attempt at introspection.

3.4 A Note on Pure Representationalism

Throughout this chapter I have argued that not only do supposed counterexamples not invalidate representationalism but they also can be dealt with in terms of pure representational properties, i.e. they can be dealt with in terms of content. In the coming chapters we will see that this strategy makes our model uniquely capable of integrating insights from Bayesian cognitive science where there are various proposals for dealing with affect, attention and even the unity of consciousness in terms of pure representational properties. The discussion of this chapter can thus be seen as a defence of the phenomenological plausibility of such views.

[17] Nagel 1971.
[18] Bayne and Chalmers 2003.
[19] Tye 2003, chapter 5.
[20] Dennett and Kinsbourne 1992.

Still there are two problems pure representationalism faces. First, it seems pure representationalism cannot explain the difference between conscious and unconscious representational states. There are a number of fields of study where unconscious representational states are frequently postulated. Psychoanalysis holds that there are repressed wishes, believes and desires. Cognitive scientists often posit deep representations that cannot be introspectively accessed. And finally, perceptual psychology seems to produce ample evidence of unconscious perception. If there are unconscious representations that roughly share content with conscious ones then the facts of consciousness cannot be wholly explained in pure representational terms.[21]

A second issue for pure representationalism is that it cannot explain the difference between beliefs and conscious representations. We said in chapter one that it is natural to conceive of beliefs as dispositional states. Dispositional states arguably cannot be conscious in the relevant sense in which perceptual states etc. can be. But now it seems that every content that can be perceived can be believed. Even if one perceives the world in terms of primitive appearance properties one can still believe that the world indeed has these appearance properties. Thus the phenomenal difference between the relevant beliefs and perceptual states cannot be explained in terms of content alone.

We will later discuss the issues involved in unconscious representation in detail. It will turn out that many aspects of unconscious representation can be explained in terms of self-representation which strictly speaking is purely representational. For the moment it suffices to note that the second problem, the problem of belief, seems to offer sufficient ground to reject pure representationalism. If pure representationalism is plausible at all it is plausible only where it is limited to occurrent mental states.

3.5 Summary

This concludes our discussion of counterexamples to representationalism. The central upshot of the discussion is that, on a liberal understanding of the contents of consciousness, we can analyze all obvious examples of phenomenal differences as differences in representational content. True, if one holds that mental representations

[21] Jackson 2003; Chalmers 2004; Kriegel 2011.

ought to be cashed out entirely in terms of representations of ordinary external world properties then these counterexamples may offer sufficient ground for rejecting representationalism, or at least pure representationalism. If we take mental states to partly be second-order representations and if we appeal to appearance properties then pure representationalism can be defended for most instances.

Let me then summarize the most important points made in part one. First, all states of consciousness are transparent in a Peircean sense. The properties one is directly aware of are properties of intentional objects. Because this is the case, phenomenal properties are representational properties. This insight sets the agenda for the coming chapters. If consciousness is representational then understanding mental representation is the key to understanding consciousness. To arrive at such an understanding will be the prime task of the coming part.

Secondly, the properties that are intrinsically represented by our sensory experiences are appearance properties or primitive ways of appearing. Not only does this view synthesize many desirable aspects of other theories of conscious content, it will also turn out to be a central to the task of explaining the puzzling aspects of consciousness naturalistically (or at least as naturalistically as possible). If we can find a plausible account of how the brain comes to represent appearance properties we will have come a long way to explaining consciousness itself.

Step Two:

Metasemantics

4 Referentialist Metasemantics

The previous chapters argued that to understand consciousness we need to understand mental representation. From the standpoint of the naturalist, this may sound like good news: Many hold that representational content is more naturalistically tractable than consciousness itself. In this chapter I will argue that the representational content of conscious states, particularly their reference to appearance properties, is precisely not explainable in terms of standard referentialist theories of representational content that try to account for representational content from the vantage point of reference primarily.

The chapter will proceed as follows. First, I will give an overview of the terminology and typology of theories of representational content. As we will see, what I call referentialist metasemantics may be considered the gold standard of naturalist thinking about representational content. However, I will then go on to argue that such theories generally face what Angela Mendelovici calls the *mismatch problem*: They radically misdescribe the properties of conscious states ascribe to their intentional objects, i.e. appearance properties. Referentialism cannot explain how it is that conscious states ascribe appearance properties to their objects. The chapter will end with a brief discussion of an alternative inferentialist approach to representational content and why this approach may avoid the mismatch problem.

4.1 How to Explain Representational Content

The study of meaning is generally called *semantics*. It is useful to differentiate however, between the study of how to think properly about the meanings of expressions, and the study of how it is that certain entities in the world come to be associated with meaning at all. We will call the former discipline *semantics* proper, while the latter shall be referred to as *metasemantics*. Thus while the semanticist wants to

have a systematic theory of complex meaningful expressions, the metasemanticist wants to know how the existence of meaningful expressions can be explained in non-semantic terms. When we try to understand how representational content is a property of brain states it is metasemantics we are concerned with primarily, though we will engage some semantic issues later on.

For our purposes, metasemantic theories can be differentiated into a number of different categories. First, we can differentiate metasemantics theories in terms of their *explanandum*, in terms of what they are trying to account for. The most obvious differentiation here is that there are *mental* metasemantic theories that try to account for the representational content of mental states on the one hand, and *linguistic* metasemantic theories, that try to account for the representational content of linguistic tokens, on the other. As we are engaged with the representational content of consciousness we will focus on mental metasemantics, thereby implicitly accepting that the representational content of the mind can be understood in relative isolation from the representational content involved in speech acts.

Following a recent trend, Chalmers has suggested a second differentiation between *first-tier* and *second-tier* theories of representational content. First-tier theories are roughly those that account for representational content of states that are not conceptually structured, like perceptions and lower level cognitive representations. Second-tier theories on the other hand engage with conceptually structured representations like sentences, thoughts and beliefs. Many hold that second-tier representational content is in some way derivative of first-tier representational content, and I will argue below that this is indeed the case.[1] Thus, we will be focused on first-tier representational content, briefly touching on second-tier content in chapter nine.

Beyond different *explanada*, theories of representational content differ in their *explanans*, what they take to explain representational content. We can differentiate roughly between *naturalist* and *non-naturalist* theories of reference, where the former take representational states to be accounted for in terms of the entities of known science while the latter hold that a proper explanation of meaning would involve the introduction of novel entities.

[1] Pautz 2021; Chalmers 2021.

It has to be admitted that the distinction is everything but sharp. At a first glance for instance, theories of representational content that posit representational content as a primitive unexplainable fact will not be naturalist in spirit. However, Adam Pautz has argued that such an account should be called naturalist as scientific theories regularly posit novel primitive entities to explain certain observations.[2] Similarly, Mendelovici and Bourget have argued that so-called *phenomenal intentionality* theories that explain representational content in terms of phenomenal properties should be considered naturalist, because consciousness is natural if anything is.[3]

I will not argue this point but circumvent it by *fiat*: For the remainder of this thesis, metasemantic theories will be naturalist if and only if they have some explanatory story to tell about the emergence of representational content from entities that fit seamlessly with our best scientific theories of the natural world. That might be an odd definition because it makes what is naturalistically acceptable relative to a certain standard of established science. But while definitions may be odd, they are never wrong. On this reading, primitivism and phenomenal intentionality theories are not naturalistic, precisely because they do not offer an explanatory story in terms of entities that fit seamlessly with the world that current scientific consensus describes.

Note by the way that my naturalism is not a metaphysical thesis about what the world ultimately consists of. Rather, it characterizes a certain attitude towards philosophical theorizing. Our naturalism is the methodological principle that philosophical theorizing should not wander too far from the treated path of science. Such a view is born from the conviction that reason, guided by nothing but logic and common sense, is quickly lead astray. That our naturalism is methodological rather than metaphysical will be of great importance in chapter nine where I will argue against metaphysical realism, a view many would associate with a kind of metaphysical naturalism, precisely on methodologically naturalist grounds.

Naturalist metasemantic projects can be further grouped into two broad classes in terms of their general explanatory strategy. *Inferentialists* try to understand representational content of states in terms of their role in inference. Naturalists will then try to give an account of inferences that is naturalistically acceptable. With regard to linguistic representational content this would involve an account

[2]Pautz 2010.
[3]Mendelovici and Bourget 2014.

of what it is for some speaker to draw an inference, while in case of mental representational content it would arguably involve giving an account of how inferences are constituted in terms of cognitive tokens or brain states.[4]

Referentialists on the other hand, take representational content to be primarily a matter of *reference*, the connection between a representation and the object it refers to. Put briefly, where inferentialists hold the relations between mental states to be the primary determinants of meaning, referentialist hold the relations between mental states and objects in the world as primary in this way. Thus, referentialist metasemantics entails identifying the reference relation with some naturalistically tractable relation like correlation, causation or similarity. Note that while there are considerable difficulties involved in trying to give a referentialist account of second-tier representational content, as conceptual representational states often refer to objects that are not part of space-time (like numbers) and those do not bear any obvious natural relation to brain states or sentences, in case of first-tier representational content referentialist metasemantics promises to give quite a straight-forward answer to the age-old question of the nature of meaning. This is why many contemporary philosophers are in fact referentialists.[5] Let's investigate their approach in a little more detail.

4.2 The Referentialist Consensus

Referentialism holds that the representational relation is identical to some naturalistically tractable relation. Plausible candidates that have been favoured by referentialists are *tracking* and *structural* relations. Let's start by considering tracking relations. The original motivation for tracking theories of representational content was the development of information theory.[6] In information theory, the information one event-type bears about another is roughly a measure of the strength of correlation between them. The idea of tracking theories of representational content is thus that correlations form the underlying substrate of representational relations.

[4] Popular defences of inferentialism are to be found in Block 1986, Harman 1987, Dummett 1991, Brandom 1996. For a critical overview, see Steinberger and Murzi 2017.

[5] Popular defences are Dretske 1986, Neander 2017, Shea 2018.

[6] See Shannon 1948. For early philosophical attempts to account for representation in terms of information, see Sayre 1976 and Dretske 1986.

There are a number of different ways to spell out what exactly a tracking relation amounts to. Dretske famously relied on informational vocabulary,[7], while others hold tracking relations to require causal connections.[8] However the detailed account of tracking may look like, it is reasonable to presume that tracking relations are natural relations. If it should turn out that they are the essential ground of representational content, this would be a great step towards its naturalization.

There is an immediate problem however. Tracking relations seem to be ubiquitous in a manner that representational relations are not. The rings of a tree bear information about its age, and the relations satisfies stronger causal constraints, however it seems problematic to say that the rings of a tree *represent* its age. Furthermore, whereas representation is a normative relation, a relation that implies the possibility of failure, tracking relations aren't normative in this sense. While some philosophers are willing to bite the bullet of holding that the rings of a tree represent its age[9], most accept that this shows that representational relations cannot just be identified with tracking relations.

Further constrains are required. The most straight-forward approach here is to introduce a *teleological* constraint that requires representational states to serve a certain function.[10] Thus, representational states will be states that *serve the function of tracking* some other state. Such approaches have to assume that there is some naturalistically innocent notion of teleological properties. These notions may either hold that teleological properties obtain in virtue of enhancing an organism's fitness in the fight for survival[11], or in virtue of contributing to self-organizational dynamics of a system.[12]

While tracking theories have dominated the naturalist discourse for many years, in recent time there has been a growing interest in revitalizing the ancient doctrine of representation by resemblance. In light of advances within the mind sciences, that have hypothesised an isomorphism between the causal structure of the environment

[7]Dretske 1986; Dretske 1995.
[8]Tye 1995.
[9]Ibid.
[10]Millikan 1989; Dretske 1995.
[11]Dretske 1995.
[12]Mossio, Saborido, and Moreno 2009.

and the wiring of the cortex, philosophers have argued that representational relations may hold in virtue of some or another notion of structural isomorphism.[13]

The idea that representational content can be explained in terms of structural isomorphism faces similar *prima facie* objections as the tracking accounts. The relevant isomorphism between neuronal wiring and elements of the environment is supposed to be *abstract*, that is, rather than pertaining to some superficial similarity (like the one between map and territory) it pertains to the structure of relations involved.[14] However, abstract isomorphisms are ubiquitous. Almost every domain of objects can be mapped onto any other domain, provided both domains contain the right amount of objects.[15] Furthermore, structural relations aren't normative: Saying that some domain is *falsely* isomorphic to some other domain seems to be a category mistake.

The tracking theorists provide a powerful template for dealing with the problem of ubiquitousness: Embed the relevant relation in a teleological context. Thus, in a similar way, structuralists typically claim that representational relations are structural relations that serve some function.[16] Thus according to structuralists, some set of objects represents some other set, if it serves some function by resembling that set.

While this is no more than a rough overview that leaves many details left to be spelled out, the general structure of naturalist referentialist theories has become clear: In the worlds of Peter Godfrey-Smith, find some naturalistically tractable *exploitable relation*[17] between the intentional object and the representational state. Then embed the relation in a teleological context that both singles out the representational relation as unique and explains its normative characteristics. The following section will argue that this kind of approach inevitably will fail in accounting for the representational content of conscious states.

[13] Gładziejewski 2016; Kiefer and Hohwy 2018.

[14] The idea here is that some domain of objects bears an abstract isomorphism to some other domain iff there exists a mapping between the two, such that if a relation holds between two objects in the first domain, some relation holds between the objects they are mapped to in the second domain. See O'Brien and Opie 2004.

[15] This is a trivial result of second-order logic, see Newman 1928.

[16] Shea 2014.

[17] Godfrey-Smith 2004.

4.3 The Generalized Mismatch Problem

Mendelovici has argued that tracking theories fail to provide an account of the content of consciousness. In this section I will argue that not only is Mendelovici's claim correct, but the problem generalizes to referentialist theories of representational content generally. Mendelovici describes the problem in the following way ("mismatch cases" are just flawed predictions of metasemantic accounts):

> In a nutshell, my argument for the claim that perceptual color representations are a mismatch case for tracking theories goes like this: Perceptual color representation represent something like surface reflectance profiles, molecular properties of objects, or dispositions to cause certain internal states in us. But this is not what they represent: instead, they represent something like primitive colors.[18]

The premises of the argument are simple. First, tracking theories predict perceptual states to represent some kind of physico-functional property. This has to be the case because otherwise it is unclear how internal representational states could bear correlational or causal connections to the relevant represented states of affairs. The second premise holds that the properties that are represented in conscious vision are primitive colors.

Mendelovici thinks that primitive colors are edenic. I argued at length that they are best construed as appearance properties. However, this point hardly touches the essence of the problem. As long as primitive colors are properties that internal representation may bear no obvious informational or causal relation to, i.e. no relation that may appear as an unproblematic explanans within a naturalist theory, the mismatch problem holds.[19] The problem as it pertains to the present account of the content of consciousness yields: Tracking theories predict the content of consciousness to be physio-functional properties, but in reality they turn out to be appearance properties. Thus tracking theories fail to capture the content of consciousness.

[18] Mendelovici 2018, p. 35-36.

[19] Note that *every* naturalist theory of representational content, including the inferentialism developed below, has to hold that there is *some* natural relation between representational states and their object insofar as they hold representation to be a natural relation. However, this is not to say that representation could appear as an explanatory primitive in a naturalist theory of content.

There is nothing special about color representation. However, intuitions regarding which states of consciousness represent their own class of *sui generis* properties differ with regard to different domains of experience. Does smell represent *sui generis* properties? Does touch? What about thoughts? My intuition is to ascent to all three questions with a falling degree of conviction. But as I guess most will intuitively agree that conscious vision purports to represent some kind of primitive properties, vision is the obvious example to work with.

Mendelovici does not address structural representations but I will now show how we can generalize the mismatch problem to cover structural representations. The same issue manifests here, though in a superficially different way. While it is not clear how there could be tracking relations between internal states and appearance properties, there is no in-principle problem involved in a structural resemblance between internal states and a space of secondary qualities, say. As structuralists are quick to realize, structuralism does not put any obvious constraints on what kind of domain could be the object of a structural representation. So why not have a structural representation of the domain of appearance properties?

This will not solve the mismatch problem. The easiest way to see this is to contemplate that, while there may be a teleologically embedded structural isomorphism between the space of all appearance properties and internal states, there still will not be a structural relation between *the properties themselves* and internal states. That is, while in some sense structural representations may represent the space of appearance properties it will represent them in so far as it has a certain structure, not in terms of the intrinsic natures of the properties involved.[20]

It will help to illustrate this point by an example. Structuralists often point to maps as prototypical examples of structural representations. Now imagine two structurally identical maps of one and the same territory. However, while the contour lines on the first map represent air pressure, the contour lines in the second represent elevation. Now it is obvious that I cannot figure out the difference between the two maps by staring at them intently. In fact, as the scenario is described, the difference cannot be inscribed on the map itself. The maps represent elevation

[20]This point is also made by Eckardt 2012 who argues that structural representations fix their referent only up to its relational properties.

and air pressure of certain points not in terms of their intrinsic properties, but only insofar as they have certain relational properties.

Now imagine for example the space of appearance properties is described by some kind of map, where neighbouring points are perceived as intuitively similar and the relevant contour lines show the intensity of emotion associated with the property. You can make this map that captures the nature of appearance properties as detailed as you like. Now it will always be possible to map the same map onto the meteorological conditions in Western Europe, say. The representation would represent appearance properties only insofar as they possess certain relational properties, not as they are in themselves.

However, it is perfectly obvious that conscious representational content does *not* represent appearance properties merely in terms of their relational properties. There is no conceivable way it could turn out that, though my perception seems to represent secondary blue, it really represents rain over London. If it can not turn out that way, then there must be an important difference in the manner our conscious states represent appearance properties and the structuralist approach. This is the mismatch problem all over again.

It may be suspected that the referentialist can provide some other plausible relation that representational content is supposed to reduce to. On closer inspection this seems unlikely. The problems elaborated above have a common structure: Figure out some naturalistically tractable exploitable relation between representational object and representational state and explain the representational relation by embedding it in a teleological context. But *there arguably is no naturalistically tractable exploitable relation between appearance properties and brain states*! This is why referentialist metasemantics predictably fail. Thus, naturalist referentialism are a dead end when it comes to providing a naturalist foundation for a representationalist theory of consciousness.

4.4 Hope for the Naturalist

At a first glance, one may think that the generalized mismatch problem deals a death blow to naturalism, as we defined it: If there is no naturalistically tractable relation between brain states and appearance properties then the relevant repre-

sentational relation can not be naturalistically accounted for. I briefly want to point out why such an argument would be rash.

Naturalists commonly accept reference of mental and linguistic states to entities to which they bear no *obvious* naturalistically tractable relation. Almost no-one would deny that we regularly refer to all kinds of abstract entities like numbers and theories. The reason some of us accept naturalism despite this fact is that we suspect that the relevant representational relations are explainable by the way we use the relevant symbolic states and by the interrelations to other symbolic states. In short, we suspect that these states have their content in terms of their interrelations in inferential processes primarily, that is, reference to these abstract objects and properties is thought to be explained through their inferential role.

The idea that I will defend for the remainder of part two is that something similar is true for first-tier representational states. In the following chapter we will see how probability theory and contemporary cognitive science makes sense of the brain as an inference machine. In the chapter after that we will approach the issue of inferentialism generally and spell out what it might mean that brain states have contents in virtue of inferential relations. Afterwards we will see that the resulting picture of the mind entails a natural role to play for representations of appearance properties within our mental economy. Finally, chapter eight will provide a complete account of consciousness in terms of inferential structure that is integrated with what we know about the neuroscience and psychological role of conscious experience.

5 Bayesian Cognitive Science

Early defenders of referentialist metasemantics were largely inspired by the application of information theory to the mind sciences and the *bottom-up* approach to cognition associated with it,[1] an approach that conceived of cognition as a passive categorization of information coming in from the environment. The naturalistic spirit of this enterprise is laudable, but as we have seen, it is doomed to failure. Naturalist referentialism will not be able to explain how we are able to represent appearance properties and thus cannot serve as the basis for a successful representationalist theory of consciousness.

The idea of treating the brain as actively engaged in constructing models of the causes of sensory data has been around at least since Hermann von Helmholtz's conception of perception as "unconscious inference".[2] Building on this legacy a new paradigm has emerged within the mind sciences that, rather than conceiving the brain as a mere bottom-up processor of information, takes it to be actively engaged with the sensory input arriving from the world in a top-down fashion. According to these theories, the brain should be conceived as constantly trying to predict future sensory input from the external world. In order to do so, it constructs probabilistic representations of the world that are updated in approximate accordance with the norms of probability theory when new evidence becomes available. I will refer to this new paradigm as *Bayesian cognitive science*.[3] A central thesis to be defended in coming chapters is that Bayesian cognitive science can be construed as the scientific backbone of an inferentialist account of representational properties, just as information theory was the back bone of referentialist metasemantics.

This chapter will provide the necessary theoretical background for all further discussion. In order to understand the Bayesian cognitive science we have to engage

[1] Sayre 1976; Dretske 1986.
[2] Helmholtz 1921.
[3] Coined in Ramstead, Kirchhoff, and Friston 2020.

two topics: Bayesianism and the cognitive science. Thus, first I will give a brief introduction into the idea of probabilities as rational degrees of belief in the light of evidence, as it is commonly understood in Bayesian cognitive science. This view is known as *objective Bayesianism*. The second part of the chapter will give an overview of how the brain is thought to processes probabilistic representations. Note that the treatment of both topics will serve as an introduction, rather than as a detailed defence of the ideas involved, which would require the space of a thesis of its own. I will then discuss the coherence between objective Bayesianism and Bayesian cognitive science. Finally, I will reflect on the view of the mind-world relation the Bayesian paradigm seems to suggest or presuppose.

The chapter will dive into some mathematical detail. For a qualitative understanding of the issues involved it should be sufficient to skim these details and focus on the discussion of the formulas. Bayes' rule is an exception to this advise and readers unfamiliar with Bayesian ideas will profit from familiarizing themselves with equation (5.2). The mathematical details are there primarily to clarify the role of generative models and predictions within the Bayesian framework, as well as the relation of inductive probabilistic inference and deductive logical inference.

5.1 Objective Bayesianism

Almost since science has started describing nature using probabilities philosophers have argued about what probability claims might mean. In this section, I will introduce the view prevalent in Bayesian cognitive science, namely that probabilities can be considered claims about rational degrees of believe in the light of evidence. According to this *objective Bayesianism*, the probability of a proposition X in the light of available information I, $P(X|I)$, is a measure of how plausible X should be taken to be in the light of I. Probability theory is rendered an idealized account of plausible inductive reasoning. Probabilities are ideally rational beliefs informed by evidence. As many Bayesian cognitive scientists consider objective Bayesianism to be the correct interpretation of probabilities in their framework, I will also rely on objective Bayesianism as a philosophical background. My own commitments here are somewhat more flexible and I suspect that Bayesian cognitive science

may be compatible with alternative interpretations of probability. Fleshing out these suspicions is tangential to the issues at hand.

In discussing Bayesianism it is helpful to differentiate between beliefs as idealized states of agents and beliefs as concrete psychological states. The application of the norms of probability theory will generally put unrealistic computational demands on agents. It demands a kind of *mathematical omniscience*, the ability to calculate probabilities to arbitrary precisions, and *logical omniscience*, the ability to know all logical consequences of one's beliefs. No real system can meet such demands. Bayesians have two ways of reacting to this problem. Either, they reconsider the rational norms they are postulating, or they give some kind of account of how actual rationality is always only an approximation of idealized rationality. We will go the latter route. That is, we will put fourth objective Bayesianism as a theory of idealized reasoning, and later explain how actual agents approximate those ideals. Where necessary, I will call the idealized states *Bayesian beliefs*.

This view of probability ought to be differentiated from superficially similar interpretations of probability. Generally, *Bayesianism* is the view that probabilities are measures of rational degrees of belief. Different schools of Bayesianism disagree on how prior probabilities are to be determined, where *prior probabilities* or just *priors* are probabilities antecedent to the appreciation and integration of new evidence. *Subjective Bayesians* hold that prior probabilities are not subject to rational norms (except they have to be normalized probability functions) and merely belief-updating has to conform to the norms of probability theory.[4]

There are some *prima facie* worries regarding whether subjective Bayesianism really captures the nature of plausible reasoning. We typically *do* hold agents accountable for their prior beliefs. For instance, independently of any evidence for what day of the year is Chinese new year, it is much more reasonable to hold that every day has a probability of $\frac{1}{365}$ rather than to hold that the probability for next Tuesday is close to 1 and the probability for every other day is close to 0. When one has no information regarding whether a certain coin is biased, it seems more

[4]For an excellent introduction to different takes on the nature of probability from the perspective of an objective Bayesian, see Williamson 2010, chapter 2. For a more neutral take, see Joyce 2004. Defences of subjective Bayesianism can be found in Ramsey 1926; Jeffrey 1990.

rational to hold that face and tail are equally likely, rather than holding that the distribution is asymmetrically biased for heads from the start.

This problem is brought to its culmination in what may be called a *statistical syllogism*. Let R be the proposition that Tom is rich. Let T be the proposition that Tom is a Texan and nine out of ten Texans are rich. It seems evident that, independently of any additional evidence, $P(R|T)$ ought to be equal to 0.9. But the subjective Bayesian would (has to) hold that this inference involves the illegitimate assumption that Tom is equally likely to be any Texan.[5]

The intuitive problems of subjective Bayesianism motivate *objective Bayesians* to hold that rational prior probabilities are uniquely determined by prior evidence after all. In order for this to make sense, objective Bayesians need to offer some way to uniquely determine probabilities in the absence of evidence. The knee-jerk way to do this is to associate all possibilities with equal probabilities. This however hinges on the ambivalent notion of "all possibilities", an issue we will return to below.[6]

As an aside, note that objective Bayesianism only assumes that there is an objective probability of any proposition X in the light of information I, $P(X|I)$. It does not assume that I necessarily needs to be true, i.e. nothing rules out conditionalizing on incorrect information.

Justifying objective Bayesianism in a bit more detail involves two main steps. First, I will argue that ideally rational degrees of belief, conditional on evidence, should be conceived as probabilities. Secondly, I will argue that priors should be as uniform as possible in the light of available evidence, or technically, they should be maximum entropy distributions. As intermediate steps I will introduce the probabilistic concepts of entropy, cross-entropy and variables.

5.1.1 Cox's Theorem

There are at least two different ways to argue for the fact that degrees of belief ought to be probabilities. Maybe the standard argument is based on the *Dutch book theorem*.

[5] Franklin 2001. For some more discussion of the problems involved in subjective Bayesianism, see Joyce 2004.

[6] Objective Bayesianism is closely connected to the *logical interpretation of probability*, according to which probabilities are generalized truth values within an inductive logic (Keynes 1921; Carnap 1950; Cox 1946; Jaynes 2003). The connection is so close that even the most prominent nominal proponent of the logical interpretation, namely Edwin Jaynes, uses the idea of an ideally rational robot to illustrate the idea, see Jaynes 2003.

The Dutch book theorem states that if an agent's degrees of belief diverge from the norms of probability theory, and betting dispositions regarding two outcomes are determined by the probability ratios of these outcomes, then the agent's expected loss will be positive. That is, if the agent plays long enough, she is sure to lose all her money. As this seems to be paradigmatically irrational behavior, it follows that to be rational entails that one's degrees of belief are determined by probability theory.[7]

Another strategy is to argue for the norms of probability deductively. *Cox's theorem* states that, given certain assumption about the notion of plausibility, it turns out that probability theory is the only consistent way of thinking about plausibility. It was originally developed by Richard Cox and made more rigorous by Jeff Paris. I will state Cox's theorem as presented in the latter source. For a rigorous proof, I advise the reader to consult the work of Paris. For a more readable but less rigorous version of the proof, consult Edwin Jayne's *The Logic of Science*.

Cox's theorem starts out by making a few assumptions about our plausibility function $f(\cdot)$:

1. We assume that the relevant propositions are all sentences that can be stated in an agent's language $\mathscr{L}$. We further assume that plausibilities of propositions in the light of other propositions have numerical values. There is a function $f(X|Y)$, where $X \in \mathscr{L}$ and $Y \in \mathscr{L}$ are propositions, where $X \wedge Y$ is consistent, and that takes values from the interval $[0, 1]$.[8]

2. We assume that logical equivalence leads to equal plausibility. If $\models X' \leftrightarrow X$ and $\models Y' \leftrightarrow Y$, then $f(X|Y) = f(X'|Y')$.

3. Logical truths are maximally plausible, logical contradictions are maximally implausible. Thus, if $\models Y \rightarrow X$, then $f(X|Y) = 1$ and $f(\neg X|Y) = 0$.

4. The plausibility of a conjunction is a continuous function of the plausibility of its conjuncts and further, conjunctions get more plausible if a conjunct becomes more plausible. Furthermore, the plausibility of the conjunction can be decomposed into the plausibility of the first conjunct, given the second is true, together with the plausibility of the second conjunct (or *vice versa*). Thus,

[7] Ramsey 1926; Jeffrey 1990; Williamson 2010.

[8] Note that we will conceive of languages as sets of sentences that can be expressed in a particular language. Thus, $X \in \mathscr{L}$ says that X is a sentence in $\mathscr{L}$.

there is a strictly increasing function c such that $f(X \wedge Y|Z) = c(f(X|Y \wedge Z), f(Y|Z))$.

5. The plausibility of a proposition is determined by the plausibility of its negation. Also, the more plausible something is, the more implausible it's negation. Thus, there is a strictly decreasing function n such that $f(\neg X|Y) = n(f(X|Y))$.

6. Finally, the last assumption roughly states that, given the right kind of evidential support, any consistent proposition becomes arbitrarily plausible. Or more exactly, for any $a, b, c \in [0, 1]$ and $\epsilon > 0$ there are propositions W, X, Y, Z such that $W \wedge X \wedge Y$ is consistent and $f(Z|W \wedge X \wedge Y)$, $f(Y|W \wedge X)$ and $f(W|X)$ are within ϵ of a, b, c respectively.

Cox's theorem states that if all six assumptions are met, then $f(X|Y)$ is a conditional probability function $P(X|Y)$.[9] In particular, $c(a, b)$ turns out to be a simple product $a \cdot b$, and $n(a)$ turns out to be $1 - a$. Thus, if the assumptions accurately describe the nature of plausibility, then it follows that probability theory is the unique logic of plausibility and rational belief.[10] Cox has thus derived a logic of rational belief from first principles.

Note that we restrict our discussion to cases where $\mathscr{L}$ is a propositional language, a language consisting of statements that can be true or false. John Williamson has suggested a framework for constructing an objective Bayesianism that relies on predicate languages, but this would needlessly complicate the issues at hand.[11]

How can Cox's and/or Paris'[12] assumptions be justified? Assumptions one and five, namely that plausibility and implausibility can be captured in terms of a single function at all will be disputed primarily by those who hold that rational attitudes are multi-dimensional. Such theorists will assume that there have to be additional values quantifying conviction or disposition to accept risks in order to make the account complete. I will dodge this kind of criticism here, as we will later see how ideas from Bayesian cognitive science can be used to construct more embracing

[9] Paris 1994, chapter 3.

[10] I use the expressions 'rational belief' and 'plausibility' as synonyms here in the sense that if there is an account of plausibility this will, by necessity, be an account of rational belief and *vice versa* (as it is rational to belief what is plausible).

[11] Williamson 2010, chapter 5.

[12] Paris notes that Cox needs to assume the sixth assumption without explicitly mentioning it.

psychologically realistic models where multi-dimensional psychological states are explained in terms of a single dimension of Bayesian belief.

Assumptions two and four demand logical omniscience from a perfectly rational reasoner. If one does not distinguish between Bayesian theory as an idealized theory of rational belief on the one hand and Bayesian theory as a theory of psychology then one[13] one may be prone to deny these assumptions. But as it is plausible that idealized rationality requires idealized computational and logical capacities, these objections do not infringe on Cox's reasoning. Thus, even if premise four may require some pondering, both assumption two and four are plausible.

Finally, assumption six may be considered the most problematic one because it essentially requires that rational degrees of belief be infinitely fine-grained. However, I would suggest that here, too, the problem is that one fails to distinguish between ideal rationality and actual common sense rationality. Plausibilities may be infinitely fine-grained even if the beliefs of actual agents cannot be.[14]

I take it that from the perspective of philosophical analysis, building a theory of probability as idealized degrees of belief on Cox's theorem is at least more aesthetically pleasing as it turns probability theory into a domain akin to logic. Edwin Jaynes comments Dutch book theorems:

> It has always seemed objectionable to some, including this writer, to base probability theory on such vulgar things as betting, expectation of profit, etc. We think that the principles of logic ought to be on a higher plane.[15]

Aesthetic preference aside, if the above defence seems unconvincing, the idea of degrees of belief as probabilities can also be defended with recourse to the Dutch book theorem.[16] An advantage of Cox's theorem, specifically with regards to our inferentialist approach to mental representation is that it illustrates the close connection between propositional logic and probability theory.

Cox's theorem seems to suggest that probability theory may be treated as an extension of propositional logic to cover continuous truth-values. But whether

[13] Such confusion may occur in some foundational Bayesian work. For instance, Ramsey holds that certain probabilistic rules are "laws of psychology" (Ramsey 1926).

[14] Additional worries are brought up in Colyvan 2003, who notes that Cox illegitimately assumes that rational belief should always be subject to classical logic. I address this assumption briefly in the appendix.

[15] Jaynes 2003, p. 425.

[16] See Williamson 2010, chapter 3.

objective Bayesianism, the view that probability theory is a theory of rational belief, and *logicism* the view that probability theory is an extension of logic, are equivalent is subject to debate.[17] The views one should hold on this will be dependent on the views one holds on the status of logic, an issue we will not discuss here. At any rate, the view that probability theory is an extension of logic may still serve as a helpful analogy. The coming chapter will argue that one of the reasons why inferentialist metasemantics has got less traction than it should have is the focus on classically logical inferences instead of probabilistic Bayesian inferences. If one thinks of probability theory as a kind of continuous generalization of classical logic, it is natural to look to probabilistic inferentialism as a generalization of "classical" inferentialism.

Let's return to our discussion of the probability calculus. Rewriting c, and considering that $X \wedge Y \leftrightarrow Y \wedge X$ we get what is called the *product-rule*:

$$P(X \wedge Y|Z) = P(X|Y \wedge Z)P(Y|Z) = P(Y|X \wedge Z)P(X|Z) \tag{5.1}$$

As a final result that will be important in calculating how the plausibilities of propositions change in light of new evidence is *Bayes' rule* which follows immediately from the product rule (5.1):

$$P(X|Y \wedge Z) = \frac{P(Y|X \wedge Z)P(X|Z)}{P(Y|Z)} \tag{5.2}$$

Some terminology. A *prior probability* or just *prior* is a probability distribution antecedently to the appreciation of some new piece of evidence. The *likelihood* is the probability of some evidence, given some hypothesis. The *model evidence* names the probability of some evidence, independently of any particular hypothesis. Finally, the *posterior probability* or just *posterior* expresses probabilities after the appreciation of new evidence. We can thus informally write Bayes' rule (5.2) as:

$$posterior = \frac{likelihood \cdot \ prior}{model \ evidence} \tag{5.3}$$

Bayesians often refer to the process of updating belief in light of some evidence as *conditionalizing* on that evidence. In our notation, the appreciation of the new

[17] Jaynes 2003 and Franklin 2001 don't seem to differentiate the two while Williamson 2010 holds that they are not equivalent.

evidence E will be equivalent to moving from $P(X|I)$ to $P(X|E \wedge I)$ for all $X \in \mathscr{L}$. Numerical values can then be calculated using Bayes' rule (5.2).

5.1.2 Propositions and Variables

So far we have introduces probabilities as functions ranging over propositions, statements that can be true of false. Later on, it will be helpful to speak of probabilities as functions of *variables*, where variables will be determinables capable of taking a variety of values.

Importantly, we can always approximate variables by propositions, using adequate priors. For instance, take the variable of room temperature t (we will use small font to refer to variables) at some point in time. Now take the propositions A_1 which yields that t is between 9 and 10 degrees, A_2 which yields that t is between 10 and 11 degrees and so on. By choosing the intervals as small as we like, we can approximate the continuous probability distribution $p(t)$[18] in terms of our propositional calculus. Naturally, the relevant probabilities will then yield $P(A_\mathrm{i}|A_\mathrm{j} \wedge I) = 0$ where $i \neq j$ and $P(A_1 \vee A_1 \vee ...|I) = 1$. I is the relevant background information.

For our current purposes, it will be convenient to work with probability distributions rather than density functions.[19] Thus we will be working with probability distributions $P(a)$ where a will take an finite number of values. I will write $P(a, b)$ for probability distributions over two variables a and b. But there is no substantial commitment involved here. One could equally use probability density functions and replace the sums below by appropriate integrations. The emphasis of finite sets of possibilities has the purpose of making the discussion more easily graspable.

5.1.3 Entropy and Cross-Entropy

In this section we will try to gain a qualitative understanding of two important information-theoretic notions that will turn out to be quite helpful below. One is

[18] Here I employ the convention that probability distributions (that can be 0 for precise values) are expressed by $p(\cdot)$, while proper probabilities are expressed by $P(\cdot)$.

[19] For a demonstration that the above take on probability can be generalized to the infinite case, see Jaynes 2003. In light of assumption six of Cox's theorem it is convenient that such a generalization is possible.

entropy, roughly a measure for the smoothness of a probability distribution. The other is *cross-entropy*, roughly a measure of the difference between probability distributions.

Information theory was developed as a rigorous mathematical foundation for the study of the capacity of information channels.[20] To understand concepts of information theory it is therefore helpful to use examples of signals that consist of symbols occurring with certain probabilities. It is important to keep in mind that the concepts are much more general than that and can be employed wherever probability distributions of any kind are involved.

Shannon entropy or just *entropy* (not to be confused with thermodynamic entropy known from physics) can be conceptualized as a basic measure of the information content of a signal consisting of symbols S_i from an alphabet $\mathscr{S}$ where each symbol is occurring with probability $P(S_i)$[21]. An intuitive measure of information content per symbol will then be the number of yes/no-questions one can be expected to ask in order to guess one unknown symbol. Imagine for instance two signals consisting of the symbols A, B, C, D. In signal one we have $P(S) = 0.25$ for all four symbols. In signal two we have $Q(A) = 0.5$, $Q(B) = Q(C) = 0.125$ and $Q(D) = 0.25$. How many yes/no-questions are required to guess a symbol?

It is not hard to see that for the first signal we have to ask *two* questions per symbol, each cutting the space of possibilities in halve. For the second signal we can reason as follows. We can first ask whether the symbol is A. Then, with a probability of 0.5, we are done. We can then ask whether the symbols is one of the pairs consisting of B and C, or whether it's D. Then with a probability of 0.25, we are done after two questions. If not, we have to ask one further question and will be done with a probability of 0.125 after three questions. This results in the following calculation:

$$0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3 = 1.75 \tag{5.4}$$

Thus, the second signal carries less information than the first: We can expect to guess a symbol from the second signal source after 1.75 yes/no-questions while we need two for the first signal source. Because the logarithm to base two of the

[20] Shannon 1948.

[21] We will later see that in our objective Bayesian paradigm unconditional probabilities are best thought of as conditional ones where what they are conditional on is suppressed. Here, the probabilities of symbols may be thought of as as being conditional on our knowledge about the relevant signal source.

number of outcomes (given by $\frac{1}{P(S)}$) gives the number of yes/no-questions to be asked given a particular symbol, we can generalize (5.4) to:

$$H(S) = \sum_{S_i \in \mathscr{S}} P(S_i) \log_2 \frac{1}{P(S_i)} \tag{5.5}$$

This is known as the *entropy* of a probability distribution. By introducing the convention that $\log x = \log_2 x$, and because $\log \frac{1}{a} = -\log a$ this is the same as to say:

$$H(S_i) = -\sum_{S_i \in \mathscr{S}} P(S) \log P(S_i) \tag{5.6}$$

Entropy is usually measured in *bits*. The number of *bits* needed to encode a message is precisely the number of yes/no question one can expect to ask in order to communicate it fully. Note that it is generally the case that smoother distributions will have higher entropy. For instance, in the above example, the first signal where every symbol is equally likely has a higher entropy than the second one. Geometrically speaking, entropy is just a measure for the smoothness of a distribution.

While entropy measures the number of *bits* needed to encode a signal, *cross-entropy* asks how different two signals $P(S_i)$ and $Q(S_i)$ are. Assuming both share a set of symbols $\mathscr{S}$ a natural measure for this turns out to be the *cross-entropy* or *Kullback-Leibler Divergence*:

$$d(P, Q) = \sum_{S_i \in \mathscr{S}} P(S_i) \log \frac{P(S_i)}{Q(S_i)} \tag{5.7}$$

Motivating this result in detail would strain the scope of this introduction. Different bases for the logarithm will result in different probability measures ($\log 2$ results in *bits*, for instance). As we will be dealing with minimization tasks primarily, the base will be irrelevant to our calculations and I will just write log.

Note that outside the domain of signal sending, the cross-entropy just measures how different two distributions are. In particular, $d(P, P) = 0$. Furthermore, $d(P, Q) \geq 0$.[22] However, the cross-entropy is not a measure of distance strictly speaking as it is not generally true that $d(P, Q) = d(Q, P)$. We will now investigate how the concepts of entropy and cross-entropy enable us to determine objective Bayesian priors.

[22] This is a consequence of Gibbs' inequality, the proof of which is beyond this introduction. See MacKay 2002.

5.1.4 Objective Priors

So far the view we have defended is clearly Bayesian, as we take probabilities to be rational degrees of belief. However, we have not differentiated our position from subjective Bayesianism which holds that probabilities are rational beliefs in the light of evidence with freely chosen priors. We have settled for an inferential rule for integrating new evidence into our pre-existing commitments, namely Bayesian conditionalization, but so far these commitments, our priors, may be arbitrary probability distributions. We will now see how maximum entropy methods may be used to constrain priors, thus resulting in an objective Bayesian framework.

For the sake of clarity, we will generally avoid writing down any probability distribution that is not conditional on something else. That is, we will avoid from now on writing statements like $P(X) = x$ wherever possible. In fact, according to the objective Bayesian, *unconditional* probabilities are typically nothing but conditional probabilities where what they are conditional on is omitted. Objective Bayesians refuse to give an interpretation to the claim that the probability of some event, a coin showing head for instance, is some definite number, independently of what information is given.

For the objective Bayesian there is no problem in talking *colloquially* about the probability of certain propositions full stop. For instance, we can say that there is an objective fact regarding the probability that a coin flip results in heads. It is natural to interpret such a probability as conditional on certain evidence. The relevant evidence will be implicit in the background information available to English speakers. For instance, background information would involve the fact that coins typically have two sides one of which is heads, together with the knowledge that typical coins are roughly fair. However, exactly what background information is relevant in a certain scenario will hugely depend on the context of utterance. The important point is that objective Bayesianism does not render common sense talk of propositions as having unconditional probabilities meaningless. Claims about unconditional probabilities can usually be interpreted as claims where the relevant evidence is implicit in the relevant background information.

So how are we to determine objective priors when we are given some arbitrary evidence? We first of all start out from a scenario where we lack all information except some logical constraint. For instance, we try to determine the probabil-

ity of a coin showing heads, knowing nothing about the coin (or maybe coins in general) but that there are two possible outcomes. The obvious answer in such a case is to settle for the uniform distribution $P_{\mathrm{U}}(X)$ in such a case. A *uniform distribution* is a distribution determined solely by constraints of symmetry. As we, by stipulation, know exactly the same about heads and tails, we know that $P_{\mathrm{U}}(H) = P_{\mathrm{U}}(T)$, where H is heads and T is tails. Furthermore, because we know that $\neg H$ entails T we know that $P_{\mathrm{U}}(H) = 1 - P_{\mathrm{U}}(T)$. Thus, $P_{\mathrm{U}}(H) = P_{\mathrm{U}}(T) = 0.5$. This is sometimes called the *principle of indifference* which may be rephrased as: Don't draw distinctions where there aren't any!

This reasoning has two great problems, one philosophical, one technical. Let's address the philosophical problem first. It seems that in assuming a uniform distribution, we have already made some assumptions about the world that cannot be justified. In particular, we have made assumptions about how to subdivide reality into precisely two different options. But this subdivision is essentially arbitrary. Just by redescribing the scenario such that H is true if H_1 is true or H_2 is true (maybe the heads-side of the coin shows a male or a female) while T isn't subdivided in this manner, we have changed $P_{\mathrm{U}}(H)$ to $\frac{2}{3}$ just by defining two arbitrary terms H_1 and H_2! This seems odd at most. Terrence Fine comments:

> If we are truly ignorant about a set of alternatives, then we are also ignorant about combinations of alternatives and about subdivisions of alternatives. However, the principle of indifference when applied to alternatives, or their combinations, or their subdivisions, yields different probability assignments.[23]

The way out of this mess is to accept, as we have already done before, that probabilities are always already relative to an agent's language $\mathscr{L}$. (Strictly speaking, we should change our notation of probability to $P(X|I)^{\mathscr{L}}$, a pedantic urge we shall suppress for the sake of readability.) One can then argue that $P_{\mathrm{U}}(X)$ should be uniform across those X that form the most fine-grained differentiation an agent can make in her language. The constraints on the uniform distributions thus aren't justified with recourse to the world, which leads one to making illegitimate

[23] Fine 1973, p. 170, quoted from Hájek 2019.

assumptions, but with recourse to an agent's representational capacities.[24] It is reasonable to assume that in our discussion of Bayes-optimal neuronal processing, the 'language' of an agent will be constituted by the most fine-grained representational capacities admitted by her representational system. Here it turns out that the problem of correct sub-divisions will not arise in practice because it is already settled by the relevant physiological constraints.

So there seems to be a reasonable way of determining a uniform distribution that makes as little assumptions about the world as possible. This solves the philosophical problem involved in the idea of choosing a maximally uniform distribution as priors. Let us now tackle the technical issue.

Of course, it will not always be the case that we have no information at all about some distribution. Imagine we impose further constraints C. For instance, assume that you are asked to estimate the probability distribution of a die where you know in advance that the distribution is biased such that the expected value over all sides S is $\sum_S P(S)S = 4.5$ instead of the usual 3.5. Obviously, in this case, the uniform distribution is a bad choice for higher values seem to be more likely that lower ones. It seems that we need some measure that tells us that a distribution is *as uniform as possible* but not more uniform.

What we need is some way of spelling out that our prior distribution $P(X|I)$ is as similar to a uniform distribution P_{U} as possible. P_{I} will be a probability distribution that satisfies I. For instance, in the example of the die P_{I} would be a distribution with an expected value of 4.5. The obvious way of doing this to search for some P_{I} that is minimally different from P_{U}! Using the cross-entropy function (5.7) as a measure of difference we can define the objective probability as:[25]

$$P(X|I) = \arg\min_{P_I} \; d(P_{\mathrm{I}}(X), P_{\mathrm{U}}(X))) \tag{5.8}$$

In words: P is such that the cross-entropy relative to P_{U} is minimal which is just a formal way of saying: Stray as minimally from the uniform distribution as is in keeping with the information you have! Evaluating (5.8), assuming a set of N propositions in $\mathscr{L}$ we get:

[24] Williamson 2010.

[25] While we defined the conditional probability before using Cox's theorem, the definition was indeterminate where I is ambiguous. We have now alleviated this flaw.

$$P(X|I) = \arg\min_{P_I} N \sum_{X \in \mathscr{L}} P_I(X) \log P_I(X) = \arg\max_{P_I} H(P_I(X)) \tag{5.9}$$

This is because $P_{\mathrm{U}}(X) = \frac{1}{N}$ is constant and can be drawn in front of the sum as a constant factor and will thus be irrelevant to the minimization. $H = -\sum_{X \in \mathscr{L}} P_I(X) \log P_I(X)$ is just the entropy. The inversion of minimum and maximum occurs because of the minus sign in H.[26]

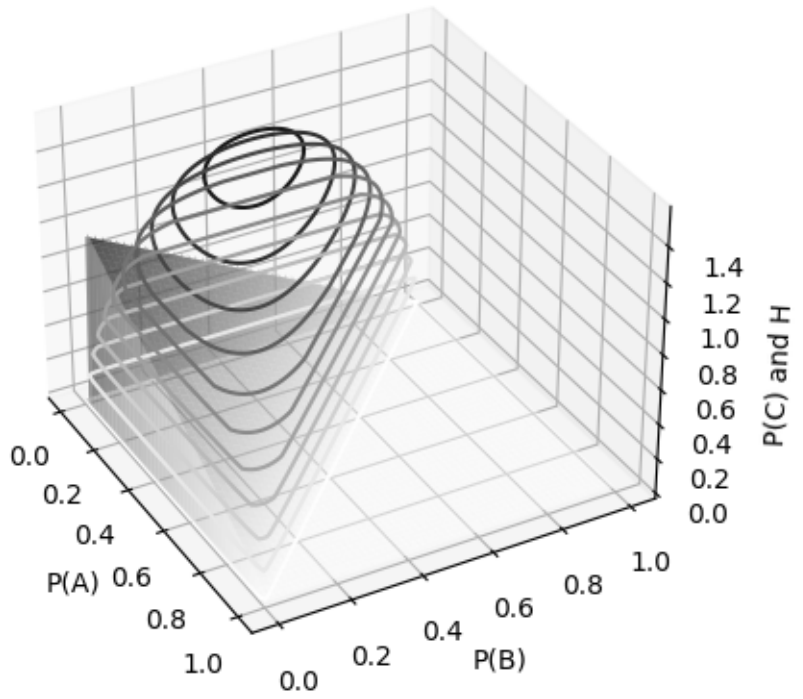

Figure 5.1: Entropy contour lines for the probability distribution over three mutually exclusive but jointly exhaustive propositions A, B and C. The probability distribution is depicted by the triangular shape, the dome above it represents the entropy. Entropy is maximal for maximally equivocal distributions, i.e. for $P(A) = P(B) = P(C) = \frac{1}{3}$ or the center of the triangular shape. This is where the contour lines reach their maximum.

What does (5.9) mean? It tells us that objective Bayesianism requires to settle for priors that possess maximal entropy. This principle is so central to the foundations of objective Bayesianism that this form of Bayesianism is sometimes called *maximum*

[26]This is the motivation of maximum entropy priors given in Williamson 2010.

entropy Bayesianism. The approach in effect requires us to choose priors that are as smooth as possible because these will be the priors that diverge minimally from uniformity. Another way of expressing the same fact is to say that reasonable priors are those that make the fewest assumptions (contain as little information in the sense of answers to yes/no-questions as possible) about the propositions in question.[27]

Back to the problem of the biased die. In order to determine $P(X|I)$, we have to maximize the entropy, subject to the constraint that $\sum_S P(S)S = 4.5$. The calculation is mathematically a bit tricky and I will not include it here. One proceeds by applying Lagrange's variational calculus which results in a distribution of the following kind: 0.054, 0.079, 0.11, 0.17, 0.24, 0.35.[28]

The philosophically important point is that we now have a principle that will help us determine unique priors, given some evidence. Thus, by extension, we now have determined $P(X|I)$ for all X and I, relative to an agent's language $\mathscr{L}$. We will now turn to the application of Bayesian principles to cognitive- and neuroscience.

5.2 Free Energy and Bayesian Cognitive Science

So far we have dealt with issues of probability in the abstract. We will now see how probabilistic approaches figure in the mind sciences. In particular, we will see how mental activity can be understood as a mechanism for approximating Bayesian conditionalization, also known as *approximate Bayesian inference*.

The *free energy principle* is a theoretical approach in the mind sciences that tries to understand the dynamics of complex systems in terms of their tendency to minimize their free energy. This may seem slightly mysterious. In the following, I will attempt to dispel the sense of mystery. It is imperative not to lose sight of the forest and get lost in the trees here. Otherwise, the general elegance of the free energy principle may be lost on the reader. So before getting into the weeds, let me give a brief overview.

Free energy is not an objective property of a system in any obvious sense. Like entropy and cross-entropy, it is a property defined over probability distributions. The free energy principle states roughly that if we describe self-maintaining systems as

[27] The method for minimizing entropy to acquire reasonable priors was originally proposed in Jaynes 1957.

[28] Seidenfeld 1986.

representing their environment in a probabilistic manner, i.e. as encoding probability distributions across states of the world, then these systems will minimize the free energy defined over these representations. For the moment we will bracket metaphysical concerns about the nature of representation and assume that there is a clear sense in which brain states may represent probability distributions over world-states. Then, saying that the brain minimizes free energy is just to say that it minimizes a certain quantity defined over its probabilistic representations of the world.

Why is this surprising? One may think that the explanation of mental life will be an irreducibly complex endeavour. Some theories presuppose that agents have at least two irreducible attitudes towards the world, belief and desire. Rational agents then are systems that act such as to maximize the desirability of the outcomes of their actions in the light of their beliefs. Other theories are still more complex, postulating additional attitudes that capture risk-aversion or conviction.

The upshot of the free energy principle can be framed as the insight that there is *a single quantity* defined over probabilistic beliefs that, when minimized, can be used to make sense of a number of, perhaps all, mental phenomena. Rather than postulating beliefs and desires, on the free energy account, rational action is understood to be the result of a single optimization process: The minimization of free energy of a single model that does not contain separate dimensions of belief and desire, just probabilities of propositions or states of the environment. The minimization of free energy not only entails plausible models of action and perception, but is also thought to underlie such vastly disparate phenomena as risk aversion, emotions, curiosity and attention.

A problem of the free energy principle is that free energy does not have an obvious intuitive interpretation. It is best dealt with as an abstract mathematical quantity. However, under certain assumptions, free energy can be expressed as *prediction error* resulting from a mismatch between predictions generated by the internal probabilistic representations and the incoming flow of sensory data. So under these assumptions, mental processes can be understood in terms of their role in a process of prediction error minimization. We will discuss this approach in more detail below. First I will introduce the mathematical skeleton of the theory and explain the concept of *active inference* that makes it possible to account for action and perception in a single formalism. Secondly, I will discuss how the free energy principle can be applied

to deal with temporality. Finally, I will introduce predictive processing to give an intuitively satisfying interpretation and neuronal implementation of the principle.

5.2.1 Active Inference

We best get to the core of the free energy principle by arguing "transcendentally" (this is sometimes called the "high road to the free energy principle"[29]). How is it that organisms acting in the world evade dissipation, i.e. what are the conditions of possibility of self-maintenance? Obviously, such systems *have to act such as to make their own existence likely*. We can imagine systems as consisting of four kinds of states. First of all, there are internal states representing environmental states. There are sensory states s that directly causally determined by the environment. There are active states a that are directly determined by the system. Then there are the environmental states e that cannot be directly observed.

We assume the system represents the environment based on some kind of implicit knowledge M about it, i.e. its prior will be $P(s, e|M)$. We call M the agent's *generative model*. The model is 'generative' insofar as it is model of how sensory states s are generated by the environmental states e.

Crucially, we will assume that M is chosen such as to probabilistically entail the agent's existence. For instance, the world-model of a fish will include the expectation to inhabit water rather than land. Thus, informally speaking, agents expect sensory evidence for their own existence. M will thus not be a model in the sense of a dispassionate image of the world, but a 'biased' representation that does not intrinsically differentiate between what is the case and what is desirable. This trick of not making an essential distinction between what is desirable and what is true is essential for describing all mental dynamics as arising from a single optimization process.

In the following, for the sake of completeness, I will give a mathematical outline of the principle. However, the most important part of this section will be the discussion following the mathematical derivations.

In light of the biased nature of the generative model we can formulate the task of self-maintenance probabilistically: The agent has to choose the actions a' from the space of possible actions such as to make the model maximally predictive. That is

[29] Parr, Pezzulo, and Friston 2022.

because we assume that the generative model is 'biased' towards desirable states. A convenient measure of the predictive power of a model is the model evidence $P(s|M)$ we introduced above, which is of course why it has that name. Model evidence measures how likely some sensory input is, given a model. Thus, high model evidence will entail that the model is very predictive and, because our model is 'biased' in the way described, this entails that the agent inhabits desirable or expected states.

The task of the organism can thus be expressed as:[30]

$$a' = \arg\max_a P(s|M, a) \tag{5.10}$$

In other words, if M probabilistically entails hat the organism survives, then increasing the chances that M is a correct model will entail survival. Saying that a model is correct means that as far as the organism can perceive, the world is such as would be expected if M were true, i.e. (5.10). The equation may thus be expressed as the imperative to act such as to fulfill a system's expectations. On the active inference paradigm, all action is conceived as an attempt to validate one's generative model. As the authors of a recent review article put it: "Given that my goals were achieved, what would have been the most probable actions that I took?"[31] It is this *prima facie* counter-intuitive view that is known as the *active inference* paradigm of action.

Evaluating model evidence is hard. Evaluating whether some sensory input is expected, given a model, involves a summation across all possible scenarios that would explain it, i.e:

$$P(s|M, a) = \sum_e P(s, e|M, a) \tag{5.11}$$

I here abbreviate sums over the range of variables x as $\sum_x \cdot$. The task therefore is equivalent to looking at all possible worlds compatible with M and sum over their probabilities. It quickly turns out that this task is intractable. There are just too many possible states the world could be in. Thus, we need some *approximate* way of calculating model evidence. As it is inessential to the following calculations we will write $P(\cdot|M)$ instead of $P(\cdot|M, a)$ to increase readability. Mathematically this is irrelevant.

[30] For a general introduction, see Friston and Ao 2012.

[31] Millidge, Tschantz, and Buckley 2021. Note that strictly speaking we are currently abstracting away from the temporal dynamics of active inference. More on this below.

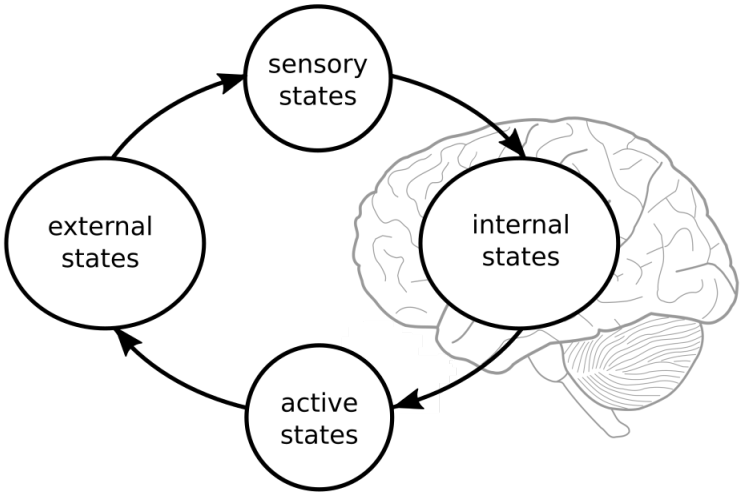

Figure 5.2: A schematic illustration of the task of all nervous systems or indeed of all life: Act such that that the world acts on you in the way you expect, given your (biased) model. Arrows indicate causal influence. In line with this figure we will refer to environmental states as e, sensory states as s and active states as a for the remainder of this thesis.

We introduce a dummy probability distribution $Q(e)$ across states of the environment that will later turn out to function as an *approximate posterior* or a *prediction*. We then define the *free energy* $\mathscr{F}$. Importantly, while the calculation of $\mathscr{F}$ still involves summing over all conceivable states of the world, we can choose $Q(e)$ such that it will be 0 for most environmental states e, significantly reducing the complexity of calculating (5.10).

While $\mathscr{F}$ does not have an intuitive interpretation, we will later see that it is, under some additional assumptions, equal to prediction error. Thus we can roughly put our approach like this. Rather than solving (5.10) directly by summation (5.11), we solve it indirectly by guessing some state of the world and then minimize the resulting prediction error, which is trivial to calculate. In effect, we evaluate model evidence by trail and error.

The mathematical expression for the free energy as a functional (a function of a function) of $Q(e)$ and $P(s, e|M)$:

$$\mathscr{F}(Q(e), P(s, e|M)) = \sum_e Q(e) \log \frac{Q(e)}{P(s, e|M)} \tag{5.12}$$

Note that $\mathscr{F}$ will be independent of e due to the summation and merely depend on s, the current sensory state. Also note that, if we multiply the model evidence (5.11) by $\frac{Q(e)}{Q(e)}$ and take the logarithm, this will not shift the location of its minimum. In the following equation, we have the logarithm of the model evidence on the left-hand side. On the right-hand side we have the free energy (because $\log \frac{1}{n} = -\log n$).

$$-\log \sum_e Q(e) \frac{P(s,e|M)}{Q(e)} \leq -\sum_E Q(e) \log \frac{P(s,e|M)}{Q(e)} \tag{5.13}$$

The truth of (5.13) is a consequence of *Jensen's inequality*, the proof of which would go beyond this introduction.[32]

(5.13) tells us that minimizing $\mathscr{F}$ will automatically fulfill the task formulated in (5.10) because $\mathscr{F}$ (right hand-side) is always greater than the logarithm of the model evidence (left hand-side)! We can say that the free energy is an *upper bound* on the model-evidence. In essence we wanted to minimize a quantity that we could not calculate explicitly and so we instead minimized another value that we know was always greater than the original target quantity. (5.13) expresses the central insight underlying the free energy principle. We have transformed our fundamental probabilistic imperative of self-maintenance, that turned out to be computationally intractable, into the imperative of minimizing free energy.

So far we have shown that the fundamental guide to action, (5.10), can be approximated by minimizing $\mathscr{F}$ over some dummy distribution $Q(e)$. We will now investigate the consequences of such an approach by investigating the properties of $\mathscr{F}$. We can decompose $\mathscr{F}$ in a number of ways. Due to the product rule (5.1), we can decompose $P(s,e|M)$ into $P(e|M)P(s|e,M)$ or into $P(s|M)P(e|s,M)$. Let's start with the second decomposition. We can then separate the logarithm and perform the sum over $Q(e)$ (because $\sum_E Q(e) = 1$) and get:

$$\mathscr{F} = -\log P(s|M) + \sum_e Q(e) \log \frac{Q(e)}{P(s,e|M)} \tag{5.14}$$

Minimizing $\mathscr{F}$ will thus necessarily result in minimizing these two values in the right-hand sum. Let's begin by giving an interpretation of the right summand, a

[32] Jensen's inequality states that for a complex function $f(x)$, where $\sum_i^n \lambda_\mathrm{i} = 1$, $f(\sum_i^n {}_\mathrm{i} x_\mathrm{i}) \leq \sum_i^n \lambda_\mathrm{i} f(x_\mathrm{i})$. For $n = 2$ this states that a straight line will always be above a convex function. From there, the inequality can be proven by induction (MacKay 2002).

term technically known as *perceptual divergence*. As you will realize, this is the cross entropy $d(Q(e), P(s, e|M))$. We said that the cross entropy is an information theory measure for difference. As $P(s, e|M)$ is a posterior, this means that when a system minimizes $\mathscr{F}$ over its representations, then it automatically minimizes the difference between the dummy distribution $Q(e)$ and the true posterior. This is why we can say that the system performs *approximate Bayesian inference*. Thus we are licensed to call $Q(e)$ the *approximate posterior* or *prediction*. When $\mathscr{F}$ is minimized $Q(e)$ automatically approximates the true posterior.

The left summand is known as *surprise*. Thus minimizing $\mathscr{F}$ entails minimizing surprise. This in effect tells us that minimizing $\mathscr{F}$ will in effect maximize model-evidence (because the logarithm doesn't shift the position of the minimum and because the minus sign turns the minimization into a maximization). Thus, minimizing $\mathscr{F}$ will entail fulfilling the existential imperative captured in (5.10). No surprise here as this is what we set out to do all along.

On a deeper level, we can say that an $\mathscr{F}$-minimizer will strike a balance between *perceptual inference*, captured in the imperative to minimize perceptual divergence, and *active inference*, by minimizing surprise. This is why the free energy principle may explain action and perception in a single analysis. We will return to the balance of action and perception when discussing our time-dependent account below.

We can give another intuitive interpretation of the principle when we decompose it using $P(s, e|M) = P(e|M)P(s|e, M)$. Again separating the logarithm and converting a division within in a logarithm into a minus sign we get:

$$\mathscr{F} = \sum_{E} Q(e) \log \frac{Q(e)}{P(e|M)} - \sum_{e} Q(e) \log P(s|e, M) \tag{5.15}$$

The left-hand summand expressed as a cross-entropy is $d(Q(e), P(e|M))$ and is commonly called *complexity*. It effectively tells us that, in minimizing $\mathscr{F}$, the approximate posterior will diverge as minimally as possible from the prior. This intuitively captures the complexity of a prediction in that predictions that diverge far from what one already knows are non-economical. Conspiracy theories are often examples of theories that offer highly complex accounts of supposedly simple data. Systems that minimize complexity of their approximate posterior will avoid

explanations of this kind. If free energy minimizing systems avoid complexity, how can we explain the lure of conspiracy theories?

The right-hand side is commonly called *accuracy* and is effectively an expected value of the likelihood, i.e. it captures explanatory adequacy. Conspiracy theories are typically theories of very high likelihood, i.e. if they were true then the observed data become very likely! Thus the lure of conspiracy theories is that they sacrifice the imperative of minimizing complexity for the imperative of maximizing accuracy. They are elaborate constructs that explain away every aspect of the available evidence.

A nice way of capturing the meaning intuitively is to say that, in minimizing $\mathscr{F}$, a system will adopt hypotheses about the world ($Q(e)$) that are as simple as possible, given what is known ($P(e|M)$), but not too simple in the sense of sacrificing predictive power. As the example of conspiracy theories seems to suggest, striking the right balance between these is epistemically crucial.

The upshot is that minimizing free energy entails an account of perception as approximate Bayesian inference, an account of action as choosing action that maximize model evidence (minimize surprise) via active inference, all the while striking a balance between making assumptions that are complex enough as to make accurate predictions possible, but no more complex. These considerations should give the reader a rough understanding of why free energy minimization is thought to be a powerful unifying principle for understanding mental activity.

Before proceeding to a discussion of time-dependent active inference I want to address an important ambiguity that pertains to the interpretation of the free energy formalism. One way of expressing its central import is to say that if we describe a self-organizing system as though it possesses a representation of its environment, i.e. a system that tends to resists decay, then this system will minimize the free energy defined over its ascribed representations. That is of course because, as we have seen, such a system has to causally facilitate self-maintenance. In this way the free energy principle captures a fundamental existential imperative. I suggest that we call such systems minimize their free energy in this sense *free energy minimizers*. Then the upshot of the free energy principle will be that all self-maintaining systems can be described as free energy minimizers.

An immediate problem of describing the free energy principle in this way is that the status of the representational vocabulary is far from clear. What justifies the attribution of representational properties, i.e. predictions and priors? Why is it valid to ascribe Bayesian beliefs to arbitrary self-maintaining systems? Questions like these have inclined a number of philosophers to adopt a kind of *instrumentalism* towards the free energy principle. These theoreticians hold that the free energy principle is best construed as a mere would-be description of actual physical dynamics.[33]

On a stronger reading, we can interpret the free energy principle more concretely as an *algorithm* that can be implemented within cognitive systems to keep entropic onslaught at bay. Under this understanding prior distributions, approximate posteriors and free energy are encoded in concrete structures within a cognitive system and the minimization of free energy is the process that describes the actual dynamics of these representations. I suggest to that we call systems that use free energy minimization as an algorithm underlying their cognition *free energy users*: They actually evaluate the free energy of their internal representations explicitly and use it as a guide (technically, perform gradient descend over it) to figure out appropriate actions. In my estimation, the conflation between free energy users and free energy minimizers has caused some confusion in debates around these issues.[34]

It is plausible that every free energy user will be a free energy minimizer. However, whether there are any free energy users at all is up for debate. That is because the thesis that certain cognitive systems like human brains are free energy users involves a substantial thesis about their functional structure and about the metaphysical status of representations. That is, the thesis that such systems are free energy users is intrinsically tied to a form of realism about representational properties where representational properties play an important function within cognition.

The following analysis will be predicated on the thesis that the human brain is actually a free energy user. In particular, the predictive processing schema to be described below is a detailed account of how the brain implements the free energy principle *as an algorithm*. Much of what I have to say is dependent upon the scientific plausibility of that very framework. Also, coming chapters will develop

[33]See for instance Hipólito, Ramstead, and Friston 2020; Es 2021.

[34]The idea of distinguishing between free energy users and minimizers emerged in discussion with Karl Friston. A similar distinction is drawn in Hipólito, Ramstead, and Friston 2020.

a realistic framework of representational properties as required by the approach. I will stay neutral on the question whether representational properties in free energy minimization should be interpreted instrumentally or realistically.

We will now discuss how to deal with predictions not of *currently* incoming sensory data but with an uncertain future.

5.2.2 Predicting Ahead

So far we have seen how the free energy principle gives an account of self-maintenance by modeling environmental states and performing approximate Bayesian inference. However, our model so far is a static one. When talking of action we considered it in the abstract, independently of a temporal dimension. Because the future-oriented predictive dynamics of active inference will turn out to be relevant for accounting for conscious experience we now have to consider how to deal with time in the framework. Also, this gives me the possibility to demonstrate a working example of a particular kind of generative model.

We will tackle the issue of time by assuming that the world we are modelling is evolving in discrete step $t_1, t_2, ..., t_N$. We also assume that our agent has a fixed time horizon determined by the number N, the number of steps it will predict ahead. We will denote the environmental, sensory and active states at time step t_x as e_x, s_x and a_x. We define what are called *policies* $\pi_i = (A_{i1}, A_{i2}, ..., A_{iN})$ as possible sequences of future action (a_i denotes an active variable, A_{ix} denotes that proposition that active state i occurs at time x). We can then conceive as the task of an active inference system by choosing the policy that maximizes model evidence.

Just as before it's practically impossible to calculate model-evidence directly. We need some method of approximation. At first one may think that we already have all the necessary tools in hand: Just see which policy entails the least free energy. But there is a caveat: We cannot straightforwardly calculate the free energy by using (5.12) because this calculation will partly depend on s, or in our case, *future values of s*! You can hardly minimize the free energy or prediction error from arising

from a sensory input that did not yet come to pass. As we cannot know these values, the vanilla free energy formalism is not up to the task.[35]

The way out of this mess is to predict future values of S and work with the *expected free energy* instead.[36] It is denoted by:

$$\mathscr{G}(\pi_i) = \sum_{e,s} Q(s_t, e_t|\pi_i) log \frac{Q(e_t|\pi_i)}{P(s_t, e_t|\pi_i, M)} \tag{5.16}$$

The central difference to the original free energy formula (5.12) is that we now sum over future sensory states, too. This enable us to calculate a *free energy estimate for expected future input*. On this natural extension of the original formalism an agent should choose policies with minimal expected free energy.

Just as before we decomposed free energy into into surprise and perceptual divergence, we can now once again employ the product rule to enable a parallel decomposition of $\mathscr{G}$. To do this, we use the fact that $P(s_t, e_t|\pi_i, M) = P(s_t|\pi_i, M)P(e_t|s_t, \pi_i, M)$ and, because model-evidence is independent of any chosen policy,[37] $P(s_t|\pi_i, M) = P(s_t|M)$. We then get:

$$\mathscr{G}(\pi_i) = -\sum_{s,e} Q(s_t, e_t|\pi_i) \log P(s_t|M) + \sum_{s,e} Q(s_t, e_t|\pi_i) \log \frac{Q(e_t|\pi_i)}{P(e_t|s_t, \pi_i, M)} \tag{5.17}$$

A brief look at (5.12) will tell us that we are just dealing with the expected values of surprise (left summand) and perceptual divergence (right summand). Tellingly, the former of these is also known as *pragmatic value* while the latter is called *epistemic value*. If we consider the 'biased' nature of M, we can say roughly that the pragmatic value of a policy scores how much an organism is expected to diverge from its optimal states, given it takes some policy. The epistemic value on the other hand quantifies the expected quality of perceptual inference, given the policy.

Taken together (5.17) gives us a clearer picture of the balance of active and perceptual inference, action and perception, we dealt with before. $\mathscr{G}$-minimizers choose paths into their future (i.e. policies) that strike a balance of having accurate

[35] Note that the difference between vanilla free energy and expected free energy can be rendered as a mere difference in our model. We merely assume a world that can be divided into temporal slices.

[36] Parr, Pezzulo, and Friston 2022.

[37] This is because the relevant information is already contained in M.

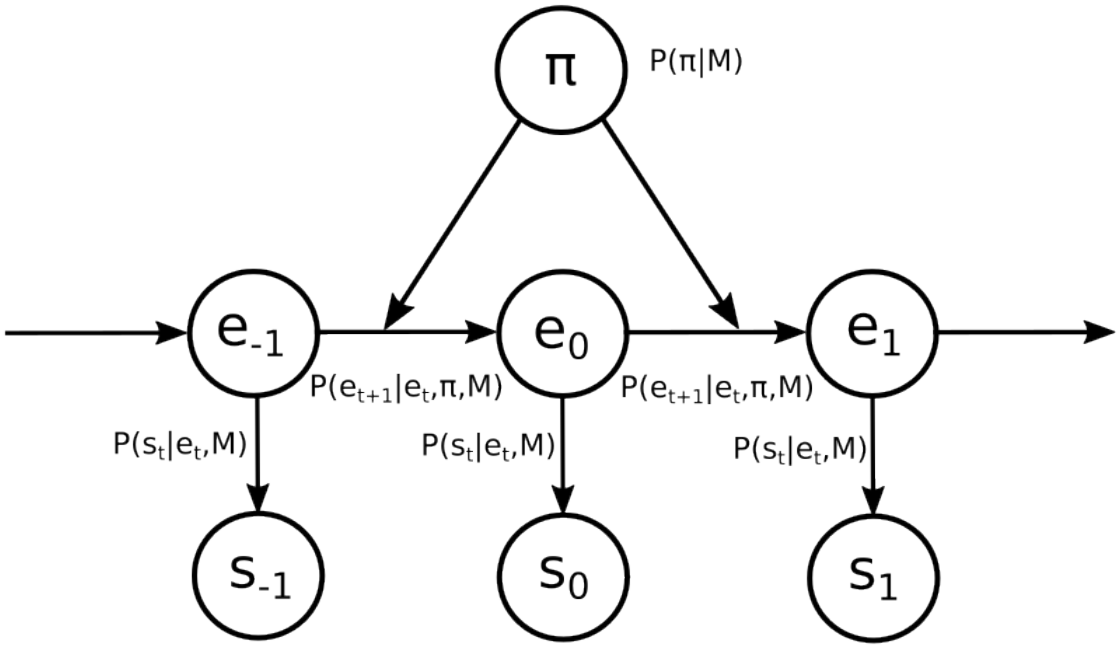

Figure 5.3: An example of a generative model as it is presupposed here, technically known as a *partially observable Markov decision process* (Parr, Pezzulo, and Friston 2022, p. 70). The illustration may also serve as an example of a simple generative model generally. Three time steps are shown that evolve from the left to the right. The leftward arrows are state transitions, encoded in the generative model as transition probabilities. The agent can act on this flow of events by the choice of a suitable policy. The states of the environment are not directly observable by the agent. All it 'sees' are the sensory consequences of the state the world is in. Note that the arrows between states can roughly be equated with represented causal relations, transition probabilities as quantifying their strength. We will make this intuition more precise later on. It is helpful to remember that internal generative models can be thought of as a set of implicit assumptions about the causal structure of the environment.

perceptions about the world and trying to reach their goals. When you try to reach a goal in an uncertain environment it is a prerequisite that you try to collect information first. But there is a balance here. You shouldn't delay action indefinitely because you try to minimize epistemic uncertainty.

A central idea of what lies ahead is that what we predict is what we perceive. If this is correct $Q(s, e)$ can, in some sense, be considered the content of perception. It is the currently best estimate regarding our environment and how we expect this environment to act on our sensory surface. For now I want to

discuss predictive processing as a neuronally plausible and intuitively satisfying instantiation of the free energy principle.

5.2.3 Predictive Processing

The much celebrated paradigm of predictive processing[38] is essentially a special case of the free energy principle. It can be derived by making additional assumptions about our model M. When it is assumed that prior probabilities consist of Gaussian normal distributions that determine the sensory states, then free energy becomes *prediction error*. The essential imperative of mental life then becomes the minimization of prediction error or expected prediction error. Here we will mainly focus on the simpler time-independent account and talk about prediction error instead of expected prediction error. Further, instead of stressing the mathematical detail[39], this section is devoted to communicating a qualitative understanding of the predictive processing framework.

Predictive processing conceives of the brain as a hierarchy of levels where every level is engaged in minimizing prediction error. Generally, every signal travelling downwards the hierarchy is a prediction. Everything traveling upwards the hierarchy is a prediction error. At the lowest level, predictions are compared against sensory information flowing in from the environment. Higher levels in the hierarchy generally represent more long-term patterns in the environment, while lower levels represent ever fluctuating sensory details. For instance, a higher level may represent the fact that you are in the garden. This context is predictive of the fact that there is a squirrel in the tree, represented at a layer below, which in turn is predictive of specific patterns of sensory input coming in at the lowest level of processing. Hypotheses are revised in reaction to high prediction error, i.e. if the expected pattern of stimulation does not actually arise. If the supposed squirrel takes of towards the sky it was probably a bird all along.

Predictive processing can be interpreted in a Bayesian fashion. The model M of the environment is encoded in the connection strengths of synapses. Specifically, higher

[38] For an overview, see Wiese and Metzinger 2017. For book-length introductions, see Hohwy 2014 and Clark 2016. For the equivalence of free energy minimization and prediction error minimization, see Friston 2005. Early forerunners of predictive processing are Rao and Ballard 1999. Finally, an early attempt to unify neuroscience under a predictive paradigm is Hawkins 2004.

[39] For an illuminating mathematical tutorial, see Bogacz 2017 and Parr, Pezzulo, and Friston 2022, chapter 5.

level activities encode priors. Downward connection strengths encode likelihoods. In order to do exact Bayesian conditionalization, we would also need to consider model-evidence, but remember that we adopted this approach to avoid calculating model-evidence directly. Thus it does not show up in our Bayesian rendering of predictive processing. In this way, predictive processing can be seen as a method of approximate Bayesian inference. It consists of making a guess and then changing the values of predictions such as to minimize the resulting error. According to the predictive processing paradigm the brain essentially performs Bayesian inference by trial and error.

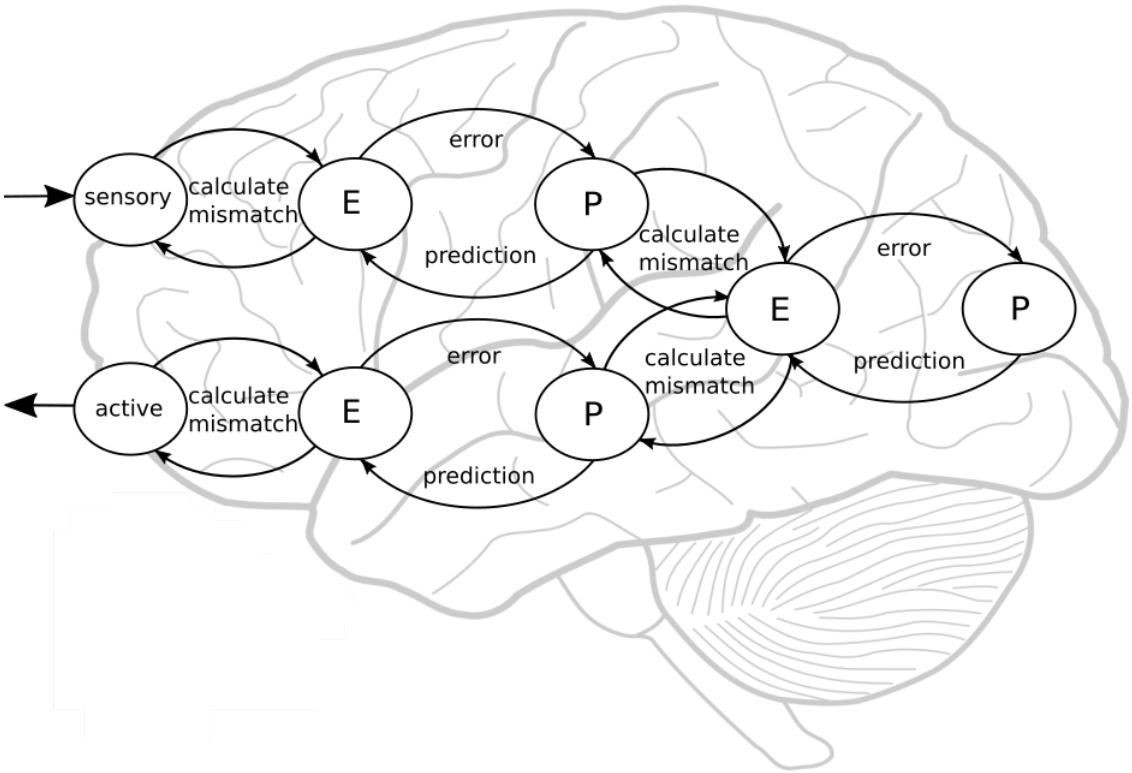

Figure 5.4: We zoomed into the fine mechanics of figure 5.2 as understood in predictive processing. This figure shows two layers of the predictive hierarchy, each consisting of a prediction unit (P) and an error unit (E). Note that in one layer two predictive units share an error unit. Ultimately, all prediction errors trace back to mismatch with sensory input. Note that from the standpoint of the system, there is no essential difference between action and perception. Both the control of active states and the prediction of sensory states result from an activity of minimizing prediction error. Note that the diagram does not show dynamics of *expected* prediction error, which would be similar in structure but more complex.

In order to compute a sensible prediction error one has to keep track of one's own predictive certainty. As an illustration, imagine you are trying to predict the weather. In order to know whether your model is good or bad, you have to make an estimate of predictive certainty. For instance, if you fail to predict rain an hour from now, this may be a reason for seriously revising your model. But if you react similarly to a failure to predict rain a week from now, this may result in a revision of a perfectly fine model! In order to make reasonable assessments of the prediction error we need a *second-order model*, a model of the quality of our first-order model.

The predictive processing schema has a natural place for a second-order model. Where the first-order model corresponds to the expected mean of a prediction, the second-order model corresponds to its expected *precision* (the inverse variance).[40] This is implemented in the predictive hierarchy as a gain on prediction error. That is, we can think of the precision as a volume control of the prediction error. Errors in predictions expected to be precise are thus regulated up, while errors in predictions expected to be imprecise are regulated down. The revision of predictions is driven by errors that were particularly unexpected.

Interestingly, this makes precision a prime candidate for accounting for the psychological property of attention.[41] Think about it this way. Items at the center of attention activate more cognitive resources because errors at the center are regulated up and drive the largest part of the revision of prediction and model. Items at the periphery activate fewer cognitive resources because errors at the periphery are largely suppressed.[42] Thus attention is conceived as a second order model.

Predictive processing also sheds some light on the nature of active inference. Above we said that there are two ways minimizing free energy, namely better models or changing sensory input. This has an intuitive interpretation in the predictive processing framework: To get less prediction error you either have to change the prediction or the incoming input. Actions are then caused by active inferences, where active inferences are basically predictions that directly cause motor behavior that is expected to reduce prediction error according to one's model.

[40] A Gaussian is fully characterized by precision and mean. They are thus all that needs to be encoded in the predictive hierarchy, given the generative model presupposed in predictive processing.

[41] Feldman and Friston 2010.

[42] For an in depth discussion of different modes of attention as they relate to the predictive processing framework, see Hohwy 2014, chapters 3 and 9.

Let us see roughly how the concepts of predictive processing relate to concepts of ordinary belief-desire psychology. Because of the hierarchical nature of the framework, predictive processing suggests that there is no absolute difference between perceptual states and thoughts. Rather, perceptions are naturally conceived as low- and mid-level predictions while thoughts are high-level predictions. Thus, perceptual representations are just low-level representations, swiftly revised in the light of changing stimulation. Thoughts are high-level representations, more context-independent and more immune to change through the flow of sensory information.

It has recently been suggested that the difference between beliefs and desires may also be mapped onto the difference between predictions (i.e. $Q(e)$) and the priors (i.e. $P(E|M)$). While it would be false to say that M incorporates no world-knowledge, it is primarily the divergence of $Q(e)$, how one currently predicts the world to be, and $P(E|M)$ that elicits actions. While I suspect that such a mapping would not be perfect, it does suggest a way how folk-psychological notions may relate to the concepts of predictive processing.[43]

A final interesting contribution to the elucidation of folk psychological notions is the suggestion that emotions may best be described as interoceptive perceptual inference, i.e. inference that tries to explain away stimulation that has its source inside the agent. For instance, pain may be thought of as a prediction that arises from a certain kind of input typically caused by bodily damage. The sense of presence and immediacy that can be or not be connected to emotions is thought to be explainable in terms of precision of the predictions involved. The imperative to act that seems to be an intrinsic part of certain emotions may then reasonably be captured by incorporating the right action expectations into the biased generative model.[44]

The integrative nature of Bayesian cognitive science and predictive processing suggests that the variety of contentful mental states are all fundamentally expressions of a more fundamental kind of representational state: Predictions. This also suggests that directions of fit are superficial characteristics rather than built into the deep structure of representational states. One may argue that the symmetry break between these kinds of states results from the dispositions of different kinds of predictions to be impacted by prediction error in different ways. States with a mind-world direction

[43]Smith, Ramstead, and Kiefer 2022.
[44]Seth and Friston 2016.

of fit are disposed to be revised when prediction error accumulates. States with a world-mind direction of fit are disposed to initialize action to cope with error.

The mapping between the theoretical vocabulary of predictive processing and the notions of folk psychology is speculative and very much a work in progress and here is not the place to argue about the details. The only commitment I will make in this regard is that the contents of consciousness are in fact approximate posteriors or predictions. More on this below. However, the integrative nature of the account makes it plausible that there will be no in-principle difficulty in explicating folk psychological notions in terms of approximate Bayesian inference. Importantly, the free energy account and the predictive processing framework as its most plausible neuronal implementation, promise to serve as a unifying framework of the mind sciences.

I will not argue the plausibility of the free energy principle, Bayesian cognitive science and predictive processing. Here, I will settle for an 'argument from authority': While there are arguments regarding the extent to which these theories explain the nature of the mind, there is no question that they are generally regarded as the most promising attempt at, and maybe the only serious contender, for providing a unified theory for the mind sciences.[45]

In the following section we will discuss the relation of objective Bayesianism and the free energy account.

5.3 Free Energy and Objective Bayesianism

In this section I will briefly point out why Bayesian cognitive scientists interpret their probabilities in an objective Bayesian manner and also point out a few ways in which the free energy principle may be used to defend Bayesianism against some objections.

The question whether the probabilities in Bayesian cognitive science are objective or subjective Bayesian beliefs is equivalent to the question whether the priors are determined by evidence or whether they are freely chosen. Objective and subjective Bayesians more or less agree on the mechanism for belief updating.

Now as we have seen that Bayesian cognitive scientists typically assume that prior probabilities are model-relative in the sense that all priors are conditional on an

[45] For an overview, see Friston 2010. For a recent critical review that essentially calls for more focused experimental efforts, see Walsh, McGovern, et al. 2020.

generative model we denoted by M. For instance, in predictive processing such models hold that the probabilities of states of the environment can be approximated by Gaussian normal distributions. For the model to determine probabilities it seems necessary to assume that probabilities are uniquely determined, given the model. Thus it seems that Bayesian cognitive science presupposes objective Bayesianism.[46]

As I have already hinted at there are other ways in which free energy minimization is a natural complement for objective Bayesianism. A central worry about the premisses of Cox's theorem was that they put unrealistic demands to the computational (i.e. logical and mathematical) capacities of agents. The predictive processing account may serve as an explicit theory of how actual psychological processes may be understood to arise from the approximation of the ideals dictated by objective Bayesianism, even where those ideals are unachievable in the limit. If such an understanding is possible, and furthermore possible under realistic assumptions about the nature of mind and brain, then the worries about objective Bayesianism based on the implausibility of mathematical and logical omniscience and the premisses of Cox's theorem turn out to be misplaced. Bayesian cognitive science offers a more rigorous grasp on what it means that psychological states arise from a dynamics of approximate Bayesian inference.

A *prima facie* reasonable criticism may be that, in his first premise, Cox just assumes, rather than proves, that there is such a thing as a plausibility of a proposition that can be expressed as a single number.[47] As mentioned, there are various multi-dimensional approaches that postulate more complex attitudes. However, it seems that the free energy principle, if viable at all, gives us good evidence to suppose that these dimensions can be expressed as mere emergent features of approximate Bayesian inference. Thus, any reasons to postulate additional dimensions of plausibility will loose much of their attraction. Note that the approximate posterior $Q(e)$ should arguably not be counted as an additional dimension, because it is just an approximation of $P(e|s, M)$ which is governed by ordinary probability theory. Thus, if we adopt the free energy account as our account of how objective Bayesian probabilities actually figure in our mental life, this makes it possible to defend all of Cox's premisses that are subject to reasonable doubt.

[46] Though this is inconclusive. If we choose a generative model that fully fixes prior probabilities, then the maximum entropy formalism would become irrelevant.

[47] Van Horn 2003; Williamson 2010

Another criticism of objective Bayesianism that may be addressed from the perspective of Bayesian cognitive science is the supposed circularity of the principle of indifference. Remember that the principle of indifferences states roughly that one should have priors that are as symmetric as possible. But 'possible' seems to be relative to a partition of possibilities. But as we are discussing inference as a process implemented in wetware, it seems plausible that our seemingly arbitrary partitioning of possibilities may be justified with recourse to an agent's representational capacities. In such cases, the generative model that is relevant will be determined by the physiological makeup of the organism. For instance, in the case of predictive processing, that different neuronal parameters encode variances and means of Gaussians is wired into the very structure of the predictive hierarchy. In such contexts, questions of the partition of reality are settled by physiological constraints. While this might not solve the problem in contexts where such constraints are irrelevant, it does seem to make objective Bayesianism a valid foundation of Bayesian cognitive science. Objective Bayesianism may not dictate principles of rationality *simpliciter*, but certainly principles of biologically constrained inference.

I would urge a kind of pluralism here. There is no strong reason for holding that this is the uniquely correct account of probability or rational deliberation. All I am committed to is that, where Bayesian cognitive science is concerned, objective Bayesianism seems to at least result in an adequate interpretation of probabilities.

In the following final section I will give the reader some heuristics pertaining to how to think about the mind-world relation in the context of Bayesian cognitive science.

5.4 The Mind-World Relation

In closing our discussion of Bayesian cognitive science and objective Bayesianism it will be helpful to illustrate the way scientists in the field think about the relation of mind and world. The view can be characterized as a form of *metaphysical realism*, a concept we will investigate in greater detail in chapter nine. The general idea is that the world is conceived as wholly independent of mental processes. One then assumes that this mind-independent world is in one of a huge number of possible

states. The epistemic task of the brain will then be to infer in which of the various possible states the world is in, based on incoming sensory data.

What allows the brain to do this is the mechanism of Bayesian inference. The interpretation of probabilities will then be *epistemic* in the sense that they represent degrees of epistemic uncertainty. They will be objective in the sense that they are wholly determined by a model and the available evidence. When new sensory states come in, the system, by using a process of trail and error, searches for an approximate posterior that would make the sensory states likely, given priors and likelihoods.

We will later see why this metaphysically realist interpretation of Bayesian cognitive science is problematic because it assumes the existence of a representational relation between internal mental states and external environmental states, rather than explain it.[48] Alternatively, I will argue that, in the light of Bayesian cognitive science, we should think of truth conditions (or 'meaning') as tied to the propensity to reduce free energy or prediction error - for a free energy minimizer a true representation is one that minimizes free energy, a false representation is one that increases free energy. The relevant concept of truth is entirely pragmatically grounded. Thus, I shall argue in chapter nine and ten, some central suppositions of metaphysics in general and the metaphysics of consciousness in particular have to be reconsidered if we take Bayesian cognitive science seriously as a unified account of cognition.

5.5 Summary and Outlook

In this section I have introduced objective Bayesianism and Bayesian cognitive science and argued that the probabilities involved in the latter are best interpreted as objective Bayesian beliefs. The Bayesian paradigm connects well to our representationalist view of consciousness. On the face of it, it offers the intriguing perspective of accounting for the totality of mental phenomena in terms of the dynamics of probabilistic mental representations. From perception and thought to attention and emotion,

[48]There are some who hold that the whole of Bayesian cognitive science, the free energy principle as well as predictive processing, should be construed in non-representational terms entirely (Es and Myin 2020). However, their arguments focus on structural representations (which I do not rely on) and an illicit metaphysical dichotomy between public and cognitive representations. In my view, there is no deep metaphysical divide here. Finally, if representationalism is true, non-representational understandings of Bayesian cognitive science will render it incapable of having anything of value to contribute to the study consciousness.

nearly every aspect of mental life has at least tentatively been integrated into the model at some point. Thus there is an obvious temptation to identify the contents involved in Bayesian theories with the contents of consciousness.

The biggest obstacle in the way of this suggestion is the afore mentioned fact that Bayesian paradigms seem to assume rather than explain representational properties. This is the central flaw we will try to alleviate in the coming chapter. In light of the demise of referentialist metasemantics, here we will instead try to explain representational properties of states by their role in approximate Bayesian inference.

6 Inferentialist Metasemantics

The *inferential role* of a linguistic or mental state is the role this state plays in inference. *Inferentialism* can be understood as a semantic, or as a metasemantic thesis. *Semantic inferentialism* identifies representational properties with inferential roles. *Metasemantic inferentialism* on the other hand holds that states have their representational properties *in virtue of* their inferential roles. For the moment we will settle for an orthodox truth-conditional semantics, that is we will assume that representational content is best captured by conditions of satisfaction as spelled out in chapter one. It is metasemantic rather than semantic inferentialism we are interested in. In this chapter we will construct an account of how mental states get associated with conditions of satisfaction in virtue of their inferential roles.

As in the previous chapter, we are presupposing a two-tiered approach to the study of representational content that differentiates first-tier basic content from second-tier conceptual content. On the face of it, as inference is a conceptual activity, inferential role semantics seems best suited for covering second-tier representational content.[1] One of the goals of this and the following chapter will be to convince the reader that this impression is misleading. In particular, Bayesian accounts of perceptual processes that describe them as unconscious Bayesian inferences warrant the application of inferentialism to first-tier representational states.

The advantage of metasemantic inferentialism will turn out to be its ability to offer a naturalistic account of representations of appearance properties. In particular, these will turn out to be low-level sensory representations within the predictive hierarchy that function as semantic primitives. It turns out that the inferential role of these low-level representations coheres with the view that they represent appearance properties. Discussing these issues however requires some preparation. This chapter

[1] Chalmers 2021.

will merely lay the foundation of our inquiry into the nature of appearance properties, which we will embark on in the coming chapter.

Our discussion will be structured as follows. First, I will elucidate the general inferentialist idea by discussing its application to the representational properties of logical vocabulary. Sections two and three will deal with the questions whether we should think of content-determining inferential roles normatively or descriptively, as internal to the subject or as stretching out into the world. Section four will discuss what meaning-conferring inferential relations are. Section five will deal with the nature of first-tier representational content. Finally, section six will discuss some objections to the emerging account of representational properties.

6.1 Inferentialism: The Very Idea

Bayesian inferentialism is a first-tier theory of sensory and sub-personal representational states, according to which representational properties of first-tier mental states obtain in virtue of their inferential role in unconscious Bayesian inference. I speak of unconscious inferences because the inferential processes involved are typically not subject to conscious awareness. This does not entail however that the contents of the relevant states cannot be conscious.

Before outlining Bayesian inferentialism, it will be helpful to gather some intuition regarding the idea of inferentialism generally. Inferentialists accounts may be divided into two very broad categories. *Linguistic* inferentialism accounts for the meaning of language tokens in terms of the rules that govern their usage in drawing inferences. *Psychological* inferentialism accounts for the meaning of psychological states in terms of the role they play in inferences, construed as psychological processes. While we are concerned here with an inferentialist take on psychological states primarily, let me start illustrating the very idea of inferentialism in a linguistic example.

The most popular example of an inferentialist account of the meaning of a term is Gentzen's account of logical junctions. Gentzen differentiates two kinds of rules that govern a term. *Introduction rules* specify the context in which a term can be introduced. For instance, for the conjunction there is just one introduction rule, namely that if A and B are part of our set of theorems, then $A \wedge B$ is also

a theorem. *Elimination rules* on the other hand, specify the moves allowed, given the term. The conjunction has two elimination rules, namely that if $A \wedge B$ is a theorem, then first A is a theorem and secondly, B is a theorem.[2]

The crucial inferentialist claim is that this is *all there is* to the meaning of $\wedge$. We may also put this as the claim that any symbol x that is used in precisely the way described, just *is* another symbol that signifies logical conjunction. I will call this kind of inferentialism *Gentzen-style inferentialism*.

While an inferentialist approach to the meaning of logical symbols may be straightforward, generalizing it to propositional contents is not. Explicating the meaning of A in terms of the valid inferences the proposition symbol is involved in will determine only its logical relations to other propositions. So for instance, if $B \vdash A$ is a valid inference, we know that if B is true, A is true. But we don't know anything about the specific matters of fact expressed by these symbols. Inferentialism of this simple kind could explain how we are able to represent the logical structure of the environment (what follows from what) but it cannot explain how we represent any specifics.

We will now turn our attention to psychological inferentialism and ask whether the idea of inferential role semantics teaches us anything about representational properties involved in unconscious perceptual inference.

6.2 What are Inferential Roles?

We first have to get clear on what exactly we mean by inferential roles. We will begin our discussion in the more familiar context of linguistic inferentialism as implied by Gentzen. A first way to understand the relevant inferential roles is *descriptively*, as descriptions of how certain linguistic entities are actually used. However, on this understanding it seems questionable whether inferentialism alone can explain linguistic meaning. For actual language users are prone to make mistakes. This however will mean that no actual language user will fully confirm to the idealized rules Gentzen identified. But if descriptive inferential roles were to determine meaning it seems that what superficially seems like a misuse of a term should actually be interpreted as a change in the meaning of a term. Thus descriptive inferentialism, according to

[2] Gentzen 1935. For an overview of the relevance of Gentzen's work for the development of inferential role semantics, see Brandom 2000.

which meaning is determined by descriptive inferential roles, faces the challenge of explaining how the distinction between right and wrong usage of a term may arise.

Thus the most natural way to conceive inferential roles is in terms of the *normative* status of the states involved, of what should and should not be inferred from them and what they should and should not be inferred from. For instance, $\wedge$ is precisely the symbol that *should* be used in accordance with certain rules. But obviously, for the naturalist, normative inferential roles are themselves in need of explanation.

A popular example of a linguistic inferentialist account of the representational content of language was developed by Brandom. In a nutshell, Brandom suggests analyzing representational properties in terms of normative inferential roles, which in turn can be explained in terms of the retributive dispositions of a language community. Oversimplifying considerably, one is using a certain term correctly, iff one is not beaten up by one's peers with a stick.[3] While this kind of account may have its merits when dealing with linguistic representations, it does not seem to generalize to the psychological case very well. *Prima facie* it is unclear what it would mean to say that some kind of inferential rule is enforced with regard to one's psychological states.

On the other hand, it is possible to appeal to biological functions in this context, just as in the context of referentialist theories. On such an account, to have a certain normative inferential role will be to have a certain causal role in a network of other states such that the whole network serves some biological function. For instance, we may imagine that some mental states of an animal perform a certain biological function in virtue of their isomorphism to Gentzen's rules of natural deduction. On this simplified account an inferentialist may hold that these mental states represent propositions and logical junctions in virtue of their normative inferential roles.

An account of this kind requires some explanation of the teleological properties such that teleological properties can in turn explain normative inferential roles. I will not follow those naturalist philosophers, mentioned in chapter four, who have tired to account for teleological properties in evolutionary terms. These hold that to have a certain biological function equals having been selected for. Such accounts entail that teleological properties are not locally supervenient on the structure of an

[3]Brandom 1996; Brandom 2000.

agent. As phenomenal content is supposed to be locally supervenient the normative inferential roles that explain this content better be locally supervenient, too.

The natural alternative to evolutionary theories of biological functions are organizational ones. These hold that teleological properties of some state are associated with that state's role in ensuring the self-maintenance of a certain encompassing system. To serve a function is to contribute to self-maintenance in the long run. On such a view, the teleological properties of states of a system are wholly supervenient on the causal organization of that system, thereby aligning with the desideratum of local supervenience.

The organizational account of teleological properties fits seamlessly with the free energy principle, as explained in the previous chapter. The entire process of approximate Bayesian inference was introduced as a mechanism for achieving organismic self-maintenance. As we showed, increases in expected free energy or expected prediction error are associated with an expected deviation from a phenotypically optimal state of an organism. Thus the systems we called free energy users will precisely be those who's representational states may plausibly be associated with organizational teleological properties.

On a sufficiently broad understanding of structuralist referentialism there is actually no difference between it and normative inferentialism. If the relevant structural isomorphism is thought to hold not between internal maps and the environment, but between the causal structure of internal states and *the inferential structure* of facts about the environment, then both views will align.[4]

In light of this alignment the question has to asked: How is it that normative inferentialism is supposed to escape the generalized mismatch problem elaborated on in chapter four, if it ultimately turns out to be a variant of structuralist referentialism? The answer is that the mismatch problem arose for structuralist referentialism because internal states do not bear any obvious naturalistically acceptable relation to appearance properties. If we accept however that appearance properties can be picked out in terms of inferential roles, which I will argue in detail in the coming chapter, then naturalist inferentialism and the particular form of structuralism that corresponds to it will be immune to the mismatch problem. In essence, while we cannot pick out

[4]Kiefer and Hohwy 2019.

appearance properties in terms of their naturalistically tractable relational properties, we can pick them out in terms of the inferential role of the states that represent them.

Before continuing, let me address an obvious challenge, namely that the account I have given so far is viciously circular. Here is the problem. Inferentialism explicates the representational content of mental states in terms of their inferential relations to other contentful mental states. As a toy example, if state a represents proposition A and b represents B and both a and b allow one to infer c, and this exhausts the inferential role of c, then, omitting elimination rules, we may say that c represents $A \vee B$. Now, while this may serve as an analysis of the content of c, it does not suffice to eliminate representational properties from the picture. We have merely analyzed the representational properties of c in terms of its normative inferential roles and the representational properties of a and b. How can we then account for the representational properties of a and b without falling into an infinite regress, never getting rid of representational properties?

Note that, for the naturalist, it is arguably not an option to hold that the apparent circularity arising here is benign. For this would entail, as we saw in chapter two, that representational properties are ontological primitives that are not explicable in terms of functional relations. If, on the other hand, representational properties can be explained in terms of physico-functional relations alone, then it should be possible to resolve the arising circularity.

A general template for dealing with issues of this kind has been provided by David Lewis.[5] The idea is that substitutions of the kind above can be repeated iteratively until representational properties are thoroughly replaced by inferential roles. We first analyze the representational properties of c in terms of inferential and representational properties of a and b, and in turn analyze the representational properties of a and b in terms of inferential roles and the representational properties of d, e and f, say. This procedure will be repeated until the representational properties of c will ultimately be replaced with an account of its place within the inferential network of states that comprise the agent's mental life. All representational notions will have been replaced by inferential ones. The resulting view holds that the representational properties

[5]Lewis 1970.

do not arise in virtue of the inferential role of any particular state on its own but in virtue of the inferential structure of a whole network of mental states.

In chapter three we saw that a representationalist explanation of unity, i.e. an explanation of phenomenal unity in terms of the unity of the intentional object, requires that the relevant underlying representations are holistic. In particular, mental representations underlying consciousness, we assumed, are such that the question of how a representational wholeness arises from representational parts does not arise. We now see that inferentialism meets this desideratum: According to this account representational states bear content only insofar as they are embedded in a network of representational states.

6.3 How far do Inferential Roles Reach?

Gilbert Harman differentiates two kinds of inferential role. *Short-armed* inferential roles are "in the head", while *long-armed* inferential roles extend out into the world.[6] For instance, the short-armed inferential role of some state representing *there is water in the glass* may include that *there is water in the glass* $\vee$ *cats can fly* can be inferred from it. The long-armed inferential roles may include that it can be inferred from the presence of the actual glass of water, out there in the world. Short-armed inferential roles connect representational states to other representational states inside the skull, while long-armed ones stretch out into the world and thereby may involve non-representational states of affairs.

As we have already noted, the representational properties of conscious experience are arguably locally supervenient on what happens inside the skull. As the relevant experiences have representational content in virtue of their phenomenal properties (even though the former may be metaphysically basic), the inferential roles that purportedly explain this content will have to be short-armed or otherwise local supervenience will be violated.

This is also in line with many defenders of Bayesian cognitive science, which asks how the brain can infer the state of the world, based on limited information available

[6] Harman 1987.

from the senses.[7] If the brain is indeed capable of constructing models of the world solely in response to the limited information implicit in patterns of sensory activity, and some assumptions hardwired into the generative model, and if this process is indeed best described in representational terms, then these representations have to be equally skull-bound. These considerations regarding the local supervenience of phenomenal content and the skull-boundness of the predictive processing mechanism in turn motivate the search for a skull-bound theory of representation. My inferentialist theory will thus appeal to short-armed inferential roles.

6.4 What are the Right Kinds of Inferences?

This brings us to what is maybe the most difficult question faced by the inferentialist: What are meaning-conferring inferential relations? What are the rules which's imposition imbue a network of states with mental content? The first candidate, that we already presupposed in our toy example above, is that meaning conferring inferential relations are those valid in terms of classical propositional logic, like the inference form A to $A \vee B$. But this would arguably be too limiting. In particular, the inferential structure underlying perceptual inference can hardly be made sense of as purely deductive inferences. Here, we need some conception of *inductive* reasoning, rather than mere deduction.

A common assumption among inferentialists is that we need a more liberal conception of inferential validity that extends beyond the notion of mere validity in virtue of form, as studied by classical logic. Such inferences are often called *material inferences*. The inference from *there is water in the glass* to *there is H_2O in the glass* may serve as an example.[8] But as this example already shows, the validity of material inferences is highly context-dependent. The inference will fail on twin

[7] This understanding of Bayesian cognitive science is defended in Hohwy 2016. For a dissenting point of view, see Clark 2017. Because of the mentioned issues of the location of representational properties of experience, I side with Hohwy. These properties demonstrate that Clark's contention that there is a reasonable boundary to be drawn around the content-conducive properties of the brain must be fallacious. I cite with Hipólito, Ramstead, and Friston 2020 that free energy minimizers can be described as representational, however when this is taken as an argument for instrumentalism about the representational across the board this threatens to level the important difference between free energy minimizers and systems with a complex internal computational architecture, i.e. free energy users.

[8] Sellars 1953; Brandom 2000.

earth, for instance. But evidently, on pain of violating the desideratum of local supervenience, the content conferring relations grounding the content of perceptual inference ought not be context-dependent in this manner.

In light of Bayesian cognitive science it is natural to hold that the right inferential roles constitutive of mental content are approximate probabilistic inferences that are characterized by free energy minimization. Thus, the idea here is that approximating the quantitative rules dictated by probability theory and the free energy principle actually are what makes some cognitive state bear its representational content.

Just as in a Gentzen-style approach certain symbolic states represent the logical interrelations of propositions, in the case of Bayesian inferentialism a set of cognitive states represents the probabilistic interrelations of a set of variables. Just as in the Gentzen-style approach the logical interrelations of propositions are represented in virtue of being guided by the right introduction and elimination rules, in Bayesian inferentialism probabilities are represented in virtue of being guided by state transitions that are isomorphic to approximate Bayesian conditionalization as characterized by free energy minimization. Note that in the Bayesian case there will be no strict correspondence to the duality of introduction and elimination rules.

In line with this argument a representational system will consist of sub-systems that each can be in a variety of states. Think for instances of neurons that can fire at a variety of rates.[9] The formalism of approximate Bayesian inference then implies that these sub-systems can be subdivided into three kinds. Those encoding priors (above denoted by the $P(s, e|M)$) and those encoding an approximate posterior or current prediction (above denoted by $Q(e)$), sensory and active states (denoted by s and a). In case of a representational system that represents its environment in terms of a time-horizon, these will have to be suitably generalized to future-directed predictions and priors. There will then exist a mapping from system states to probabilities such that the dynamics of system states track the probabilities characterized by the free energy principle in virtue of the system's intrinsic causal structure. The system will then change its prior-encoding and its prediction-encoding states under the constant flow of sensory input. Finally, Bayesian inferentialism will hold that Bayesian agents are capable of performing active inference. That

[9]The formalism is neutral on what exactly a sub-state is. One could also implement a system such that a whole cluster of neurons constitutes a single sub-state.

is, the active states will will be suitably determined by the dynamics of prediction encoding states. There may be a fourth mode of inferential activity associated with learning, i.e. a change on prior encoding states.

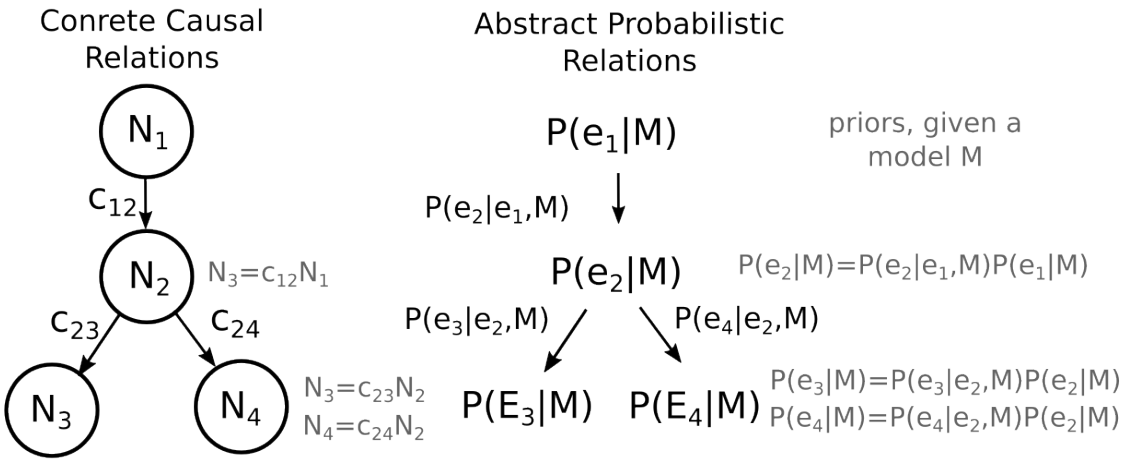

Figure 6.1: Schematic illustration of an isomorphism between causal and probabilistic structure.

Schematic illustration of an isomorphism between the causal structure of a concrete system and the probabilistic structure of a set of probabilities, relative to a generative model M. On the left you see an system consisting of four sub-systems N_i. These sub-systems can then be in a variety if states, here symbolized by numeric values associated with N_i. The depicted sub-systems are causally connected in a specific way that ensures that changes in the values of higher-level states causes change at lower levels (think of hierarchically connected neurons and strengths of synaptic connections). The precise impact of state N_i on N_j is quantified by c_{ij}. On the right you see a network of probabilities of propositions. On the highest level priors, given model M, are encoded. Causal and abstract rules run parallel. While there is some similarity to a predictive processing system, as it stands the depicted system does not involve a mechanism for minimizing its free energy or prediction error and is thus not sufficient to possess representational properties on Bayesian inferentialism. To do this one would have to include recursive upward connections signifying prediction error or implement another mechanism for calculating an approximate posterior. The figure merely illustrates the idea of an isomorphism between an abstract and a computational structure. Note that this may also serve as an illustration of the idea that a generative model may be implicitly encoded in neuronal wiring.

An immediate issue of the approach will be that, just as Gentzen-style inferentialism will merely fix the content of symbols up to their logical interrelations, the Bayesian approach seems to fix content of cognitive states only up to their probabilistic interrelations. That is, even if some representational state is veridical, all that is known about the world will be that certain state is probable and that it makes some

other states more or less probable. This is a serious worry and it will be dealt with in much more detail in the following discussion. For the moment it should suffice to say that Bayesian inferentialism differs from mere Gentzen-style inferentialism in that sensory states are involved: Where Gentzen-style inferentialism describes the dynamics of abstract *deductive* systems, Bayesian inferentialism describes *inductive* dynamics, sensitive to information flowing in from the environment. It is this fact that imbues internal representations with world-directed representational content.

Technically, Bayesian inferentialism will be a variety of *computationalism*, i.e. it holds that systems have representational properties in virtue of their computational properties (together with teleological properties). To characterize a system in terms of its computational properties is to characterize it in terms of the algorithms or formally conceived state transitions of its internal states. Within the philosophy of computation there are a variety of accounts that differ slightly in their precise take on what it means for a concrete physical system to implement formal state-transitions. For instance, *simple mapping accounts of computation* involve the mere existence of a suitable mapping between the physical states and computational states such as to preserve computational relations, as described above.[10] But it may be required to put in place more complex constraints on mapping such as that singular physical states are mapped to singular formal states in such a way that causal-dispositional structure mirrors formal structure. These issues however are tangential to the task at hand. I here merely require that there be some non-trivial notion of computational properties.[11]

Note that there is no reason to restrict representational properties to systems that use a prediction error minimization to approximate Bayesian inference. All that is needed is that the system encodes an approximate posterior, priors, sensory and active states such that the system's dynamics conforms to the principle of free energy minimization. There is evidence that the brain is using a prediction error minimization formalism and thus that this is the way the brain approximates Bayesian inference. The idea here is that this is the particular way that natural selection implemented the more general computational architecture of approximate

[10] Chalmers 1996; Chalmers 2012c.

[11] See for instance Godfrey-Smith 2009, Piccinini 2007. For an interesting reply to the ingenious anti-computationalist argument in Maudlin 1989, see Klein 2008.

Bayesian inference in the human brain, but not the *conditio sine qua non* of being a representational system. Free energy minimization is the constitutive norm of representational systems. Predictive processing supposedly is the particular mechanism that implements approximate Bayesian inference in the brain.

Finally, there is the question of *how approximate* Bayesian inference needs to be for representational properties to obtain. *Not* doing Bayesian inference at all is just the limiting case of approximating Bayesian inference. In order for our account not to be inflationary, we need some more precise way of spelling out the notion of approximation involved. An obvious intuitive measure for this will be the cross-entropy between the priors and the approximate posterior. Alternatively, one may hold that whenever Free energy is minimized representational properties obtain. Or, as a third option, one may hold that the temporal depth of processing, i.e. the degree to which *expected* free energy is minimized is relevant.

However, my intuition tells me that there arguably is no clear line to be drawn between the representational and the non-representational. Some systems possess the right causal structure to possess complex representational properties (like brains), some systems are too simple to possess any (like grains of sand), and a lot of systems may inhabit the ontological borderlands in-between (like bacteria and plants). Authors like Daniel Dennett have long argued for such a vagueness of the representational[12] and I do not intend to offer any hard criteria myself. On my view, representational properties emerge gradually when the computational-causal structure of a system becomes more complex and acquires lower degrees of perceptual divergence and higher degrees of free energy minimization and temporal processing-depth.

Some may claim that in combination with representationalism, the alleged vagueness of representational properties may become problematic. For, it is often claimed that we cannot make sense of the notion that phenomenal properties are vague.[13] I don't share these intuitions, but if one is moved by such considerations one would have to search for a non-arbitrary cut-off point.

A nice consequence of what we have learned so far is that we can draw a more principled distinction between free energy minimizers and free energy users, i.e. mere self-organizing processes and proto-cognitive activities. The difference will be one of in-

[12]Dennett 2017.
[13]Searle 1994; Schwitzgebel 2020.

ternal organization, particularly one of computational properties. In particular, a free energy user will possess internal states that are causally isomorphic to a set of probabilities, and these internal states will play a constitutive role in the self-maintaining dynamics of the system. No such constraints apply to free energy minimizers.

6.5 Representational Content

We now have some conception of how representational properties arise in virtue of their role in perceptual and active inference. What we have not tackled so far is the question of what the representational content of the relevant internal states will be. In predictive processing for instance, it is common place to say that certain neurons in the hierarchy represent particular hypothesis about the external world. What are these hypotheses?

A good heuristic for thinking about conditions of satisfaction in the Bayesian paradigm is to think about a state's propensity to reduce prediction error. It is natural to assume that a certain state is assigned the value 'true' in the predictive processing schema if it has the propensity to reduce prediction error.[14] A state is 'false' if it has the propensity to entail high prediction error. It would be unclear how a prediction could be true or false in the predictive processing schema, if it did not somehow impact prediction error under the right circumstances.

This would imply a simple semantics for predictive processing. Hypotheses-encoding states modify predictions sent downward in the predictive hierarchy. Ultimately, these predictions are matched against sensory input. If the predicted sensory inputs come about the prediction was correct, otherwise it was not. The truth-conditions of predictions are thus straightforward. They are given by the right modifications of expected future sensory input at the lowest level of the hierarchy. Objects and states of affairs are then represented in terms of their expected impact on future sensory input. Elements higher in the hierarchy encode more complex modifications of this kind, elements further down encode sensory detail.

In this raw form our semantics for predictive processing is too simple. For it seems that typical perceptual illusions are produced precisely because the predic-

[14]Hohwy 2014, p. 174.

tive hierarchy adopts a certain hypothesis that *currently* explains away incoming prediction error, but is an illusion none the less. For instance, in the Müller-Lyre illusion[15] two lines are perceived different in length because of misleading context clous. However, the current account seems to tell us that the perceptual state is actually correct as it minimizes prediction error. For this is arguably how the brain settled for this perceptual hypothesis in the first place.

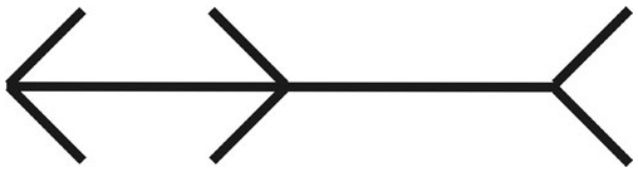

Figure 6.2: Müller-Lyre Illusion.
The Müller-Lyre Illusion. The two lines seem different in length even though they are equal. One can easily see this by covering the middle arrow head with your fingers.

To alleviate this flaw we should think of truth as an idealized reduction of future prediction error. The Müller-Lyre illusions can be dispelled by intervening on the image. By covering the middle lines with one's fingers the distortion vanishes. One then sees that both sides are in fact equally long and the original prediction assigned a high prediction error by the perceptual system. We will cover more troubling sceptical cases in part three. For now this superficial treatment will be enough to show that our account can deal with illusions as short-term error minimizing hypotheses. A state will thus be true if it reduces prediction error in the long run, false if not.

Abstracting away from predictive processing and addressing the issue of representational content more broadly, we can say that representational states represent states of affairs as modifying probabilities of future sensory input, i.e. as making certain sensory events more or less probable in the future. For the moment we will take the fact that sensory states, which are strictly speaking representations themselves, just represent "sensory input" or the direct causal influence of the environment. This view will be suitably refined in the coming chapter where I will argue that low-level representations represent appearance properties.

[15]The image is from Gentaz et al. 2004, thanks to Edouard Gentaz for permission to use it.

Two corollaries follow directly from this analysis. First, representations are best thought of as possessing continuous truth-values.[16] This helps to make sense of the fact that hypotheses in the predictive brain are represented probabilistically and are typically not revised in an all-or-none fashion. Incoming sensory data are matched against hypotheses with more or less success, entailing minor or radical revisions. Accordingly one may classify hypotheses about the world as more or less correct, depending on how well they match future sensory input.

Secondly, conceiving as truth-conditions as mere conditions of minimizing prediction error entails a form of *semantic holism*.[17] Semantic holism is the view that a set of representations do not determine truth-conditions when considered separately, but only when considered as a whole. In relation to the predictive hierarchy, a single state, like the state of a particular neuron, will determine truth-conditions only in relation to all other elements of the hierarchy. For all these elements impact the way the particular states will predict lower-level activity.

Holism is often viewed as a threat to the explanatory power of inferentialism. If content obtains in virtue of inferential role, then there are two views one may take on how two representational states may bear identical content. Either, content-identity requires complete identity of inferential roles. But then it seems that two agents can only ever represent the same state of affairs if they share the entire web spanned by the inferential relations of their representations. To avoid this, an inferentialist would have to restrict the relevant inferential roles to some special class of content-conferring inferential relations. Solutions of this kind may be called *molecular* inferentialism. However, the distinction between content-conferring and other inferential relations is quite close to the distinction between synthetic and analytic truths Quine attacked, and according to many philosophers conclusively refuted, in his paper on the two dogmas of empiricism.[18]

While objections based on holism may be interesting in the context of second-tier representations, they arguably hardly matter from the perspective of accounting for first-tier content. For here the inferentialist can arguably just bite the bullet and

[16] Kiefer and Hohwy 2018.

[17] The holism is semantic, rather than metasemantic, because it pertains to the representational content rather than to the features grounding that content, as was the case in our discussion of the interdependence of representational properties which we solved by Ramseyfication.

[18] Quine 1951. Formulated as an argument against inferentialism, see Lepore and Fodor 1993.

accept that content is holistic in the sense implied. In the context of second-tier content such an admission seems to threaten the possibility of communicating the content of thought. But in the context of perceptual and sensory contents such a consequence is much less daunting. Furthermore, I will later argue that a meaningful conception of shared content can be recovered even if one accepts holism.

An intuitive objection to our inferentialist account of representational content is that it does not offer a basis for substantial mental representations. In particular, it will end up representing the world in terms of probabilistic relations between states of affairs rather than their causal structure. For instance, one internal representation may represent a probability distribution over some variable a like air-pressure. Technically, this content will be cashed out in in terms of probabilistic relations to the right sensory input, but we will ignore this fact for now. We may further assume that some other variable b pertains to the intensity of rain. Furthermore, there is a probabilistic relation between a and b, i.e. $P(a|b)$ is such that low air pressure makes rain more probable.

Such a representation cannot purport to capture the *causal* structure of the environment. It does not distinguish between scenarios where low air-pressure causes rain, where rain causes low air-pressure and where there is some *confounding cause* c that causes them both![19] The representation of the environment in terms of probabilistic structure seems to lack causal substance.

Luckily, this kind of flaw can be alleviated by considering active inference. Let's call the variables directly determined by the agent *active variables* and let us for the moment assume that they are just range over basic behavioral facts, like moving one's limbs. Representational content will then be further grounded in their probabilistic relations to active variables. Suppose for instance, that in the above example of correlation, there was an active variable *act* that a probabilistically depends upon. Now it seems that the two of the three cases described above suddenly become decidable. If the correlation between a and b is causal, then an intervention on a via *act* will be followed by a change in b. If the correlation is the result of a confounding cause c, this will not happen. (See figure 6.3.)

[19] Pearl 2001.

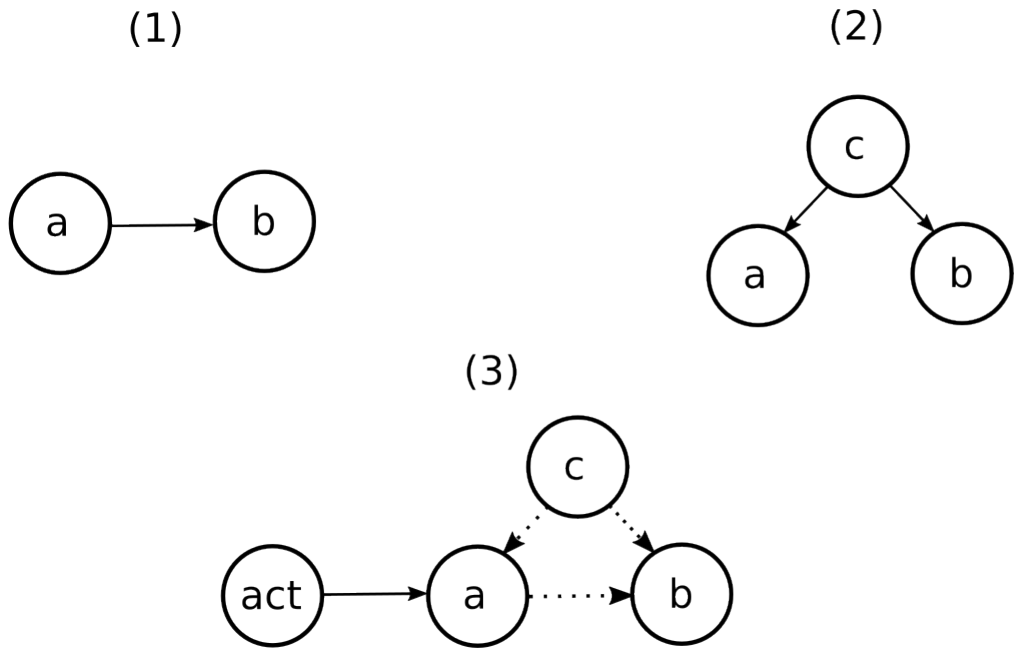

Figure 6.3: A schematic illustration of causal relations among variables.
A schematic illustration of causal relations among variables. Arrows represent causal connections. Just by observing a and b it will be impossible to differentiate (1) and (2). c is called a *confounding* variable. But if we *act* on a, specifically if we introduce an active variable *act* that takes any value we choose, we can differentiate between the cases: If causing a increases b, then the relation wasn't created by a confound and (1) is the correct diagram. If it doesn't, then c was the real cause of b and (2) is the correct diagram. In a nutshell, that's why including active variables like *act* makes it possible to represent causal affairs by representing probabilistic relations.

The upshot of this is that, by integrating active inference into the picture, Bayesian inferentialism can explain how systems represent the causal structure of the environment. As a special case, agents will typically not merely represent the probabilistic but the causal structure of their environment by representing them as a place for action. The analysis given here of course presupposes an *interventionist* take on causality, i.e. the view that causal facts are supervenient on facts about the results of interventions on systems.[20]

Interventionism also helps us fend off another objection. Just as in Gentzen-style inferentialism representational states only ever represent propositions in terms of their logical relations, Bayesian inferentialism only ever represents propositions in terms of their probabilistic interrelations and their probabilistic relations to low-level

[20] Woodward 2003.

sensory and active states. On an intuitive level it may seem that this feature makes the representational purport of Bayesian inferentialism rather weak. However, if causal relations supervene on probabilistic ones, then the systems described by Bayesian representationalism will represent their environment in terms of causal structure. It is a common view that perceptual states represent the causal structure of the environment that impinges on the sense organs. On the presented view the inferentialist-interventionist analysis of perceptual content captures this intuitive view. We will return to it in more detail in the coming chapter.

Note that the picture of representational content I painted in this section seems to undercut the metaphysical realism I argued is implicit in the principles of Bayesian cognitive science. If to be true is ultimately grounded in the minimization of prediction error and the whole process can be cast as a process of action-guidance, then it is hard to see how we could ever as much as refer to a putative mind-independent reality. The true description of the world will, from the perspective of the predictive mind, just be the description that is maximally suited for the guidance of self-maintenance and there can be no "transcendent" truth that is not somehow pragmatically grounded. We will follow these considerations in detail in chapter nine. For now I will bracket them.

6.6 Objections

This section will discuss some objections to Bayesian inferentialism. First, we will ask whether Bayesian inferentialism has a story to tell about the explanatory role of representational properties in cognitive science. Secondly, we will discuss whether cases of irrationality offer a counterexample to the supposed constitutive role of Bayesian principles in the metaphysics of content. Finally, we will ask whether computational theories of mind can be refuted generally by appealing to intuition.

6.6.1 Why are Representational Properties Interesting?

Be it in the context of Bayesian theories or beyond, representational properties are central to explanatory strategies employed in orthodox cognitive science. A theory of representational properties ought to have something to say about why representational properties are useful in this manner. It may not be entirely clear

however how such an account may be derived from Bayesian inferentialism. In other words, it may not be entirely clear what merits the description of systems in the representational vocabulary of predictions, hypotheses and errors brings.

To get a grasp on the issue it is helpful to first consider the standard way of dealing with representations in cognitive science. Very briefly, representational states, according to referentialism, serve as various kinds of stand-ins for external states of affairs, thereby allowing for complex behavior inexpiable (or at least involving very unnatural epicycles) without them. If, for instance, a rat has a map-like structure in its brain that encodes the structure of a maze, this insight will be profoundly helpful in explaining how it found its way to the cheese.

Importantly, Bayesian inferentialism can retain most of the explanatory advantages of representational orthodoxy. While representational properties are not understood as stand-in relations to external states of affairs (as the referentialist assumes), they can be understood as *a priori constraints on stand-in relations*. That is, if probability theory dictates *a priori* norms of plausibility, then any system performing operations over representational stand-ins in a noisy environment will have to work in accordance with probability theory. Thus, we can always describe a system performing approximate Bayesian inference as though it had stand-ins for external state of affairs in the was the referentialist holds. Referentialism and inferentialism here do not differ in explanatory power but merely in their metaphysical view regarding the question in virtue of what representational properties arise.

Again consider the example of a map somehow encoded in the rat's brain. Probability theory tells us that if the rat gathers new evidence about its environment, say that there is a barrier where there was none before, given that the sensory input is noisy (i.e. probabilistic), the updating of its internal map will necessarily involve some way of approximating Bayesian inference. In the same way, if the rat has to choose the correct path based on its internal map, active inference regarding future states will be involved.

This means that following the norms of probability theory is what it is for the rat to, loosely speaking, *treat its internal states as a map*. Treating internal states as a map involves the accord with two kinds of rules. First, the map has to be made responsive to the right kind of evidence coming in. Secondly, other states

have to be responsive to the representational import of the map, i.e. the map has to inform further cognitive processing in the right manner. Taking something as a map therefore, is arguably determined by descriptive inferential role in approximate Bayesian inference that involves perceptual and active inference.

These are delicate issues and they don't take center stage in our current discussion. However, insofar as it is a reasonable desideratum that a theory of content should have something illuminating to say about the role of content in cognitive science, it was important to show that such an account is possible. It is possible because inferentialism inherits some of the explanatory power of classical referentialism. Spelling in out in detail is an issue for another time.

6.6.2 Irrationality

Bayesianism is, among other things, an account of ideal rationality. Bayesian inferentialism may be roughly framed as the claim that representational properties are determined by certain constitutive norms of rational probabilistic reasoning. Accounts of this kind face an obvious intuitive reply: What about cases where agents having representational states violate these norms? How are cases of irrationality possible if certain rational norms of inference are constitutive of the very existence of representational properties?!

In our current paradigm we may want to distinguish three kinds of rationality and correlated cases of irrationality. First, an agent may be rational in a Bayesian sense but irrational in a *social* sense. A member of a cult, we may suppose, may base her believe that the world is coming to an end soon on rational Bayesian belief updating. Such a person, absent any sophisticated knowledge about the world, may have a strong prior to trust her peers who all tell her they know that the world is going to end. Still, we would intuitively say that there is a sense in which the belief that the world is coming to an end is irrational.

Cases of social irrationality are not in any way an obstacle to the ascription of representational properties according to Bayesian inferentialism. The agent, we supposed, follows all constitutive norms correctly and therefore is subject to ordinary representational explanation.

A second sense of irrationality is that an agent may follow the norms of Bayesian reasoning only approximately and is therefore *weakly probabilistically irrational*. As, according to Bayesian cognitive science, we are all examples of weak probabilistic irrationality, such a case also will not count as a problem case for our account either.

So the only supposed problem cases would be cases of *strong probabilistic irrationality*, cases where the norms of probabilistic reasoning are approximated only very grossly, or violated entirely. In these cases however, the question is open whether the description of the relevant activity as a case of irrationality, that is as a violation of rational norms, rather than *arationality*, the inapplicability of rational evaluation, is warranted.

Consider a stereotypical example of mental illness. Hans sees a bird in the sky and thinks it is the UFO that comes to pick him up. Can we describe this irrational mental state as a strong breakdown of probabilistic rationality? Arguably not. Hans still engages in approximate Bayesian belief updating, based on some quirky prior beliefs. Furthermore, Hans will probably be worried, or, depending on his expectations regarding the UFO, happy, that the UFO arrives. His mind will thus draw appropriate inferences from his perceptual experience, based on his priors. It is plausible that, if none of the inferences are correctly drawn, then the mental states of Hans can no longer be described representationally.

This, I suggest, is the solution to the puzzle of irrationality and constitutive rational norms of inference. Most kinds of irrationality need representational properties to be applicable at all and can thus only be ascribed against a background of the approximately correct application of Bayesian norms. Where these rules aren't followed even approximately, metasemantic purport breaks down and representational capacities fade away. The lesson is that strong probabilistic irrationality arguably ought not be called a case of irrationality proper because the relevant rational norms do not apply. They are arational rather than irrational.

6.6.3 The Ploy of Funny Instantiation

As a final objection, we will discuss attacks on computationalism that involve general worries that causal pattern aren't enough to instantiate mental, and

particularly phenomenal properties. I will argue that such attacks involve an over-reliance on human intuition.[21]

Many attacks on computationalism rely on thought experiments that try to point out supposedly absurd consequences of the view. If computationalism were true wouldn't this imply that a mechanism made of beer cans[22], a construction of water pipes operated by John Searle[23], the Chinese nation communicating with walkie-talkies[24] or even the USA in its current form[25] would need to be conscious? How absurd! And thus it is concluded that computationalism must be absurd, too. Tim Maudlin fittingly calls this line of argument the *ploy of funny instantiation.*[26][27]

It may be asked whether Bayesian inferentialism (plus representationalism) is vulnerable to the ploy at all. For, according to the view, the computational properties of a system are insufficient to fix mental content. Teleological constraints are necessary as well. However, while I have not worked out an explicit theory of teleological properties, I have committed to a broadly organizational account. It seems that organizational teleological constraints on mental content will not offer sufficient constraints on mental content to avoid the ploy, for arguably organizational teleological properties may obtain in virtue of causal structure alone, without reference to an underlying medium.

A good example here would be the scientific discipline of *artificial life*. The discipline focuses on the simulation of simple life-like processes that may have the properties of self-maintaining within their simulated environments. Thus these

[21] Some of the discussed replies are leveled against *functionalism*, the view that causal interrelations determine mental properties, rather than computationalism. I find it hard to distinguish between these views in a precise way, however I suspect that computational properties are best conceived as fine-grained functional properties. Thus, all *a priori* attacks on functionalism will be attacks on computationalism, too. I will thus deal with charges of both kinds in a defence of computationalism.

[22] Searle 1984.

[23] Searle 1980.

[24] Block 1978.

[25] Schwitzgebel 2015. To Schwitzgebel's credit he tentatively endorses the view that this might be true.

[26] Maudlin 1989.

[27] Another influential objection claims that functionalism is trivial in the sense that any sufficiently complex physical system has all possible functional properties. (Putnam 1987, appendix, Searle 1994, for a reasoning that is similar in spirit, but less ambitious, see Block 1978) Refuting this charge is tedious (a good account can be found in Chalmers 1996), however I think there is a good *prima facie* reason to be highly sceptical of this conclusion. A system that were to have all functional properties would in effect process an infinite amount of information in the sense of Shannon. However, it is a well understood fact from the thermodynamics of computation that such a feat is arguably impossible. In fact there is probably a limited amount of information available in the cosmos. Thus, I do not deem it necessary to discuss these results in detail here.

processes may be associated with organizational teleological properties just in virtue of their computational structure. A defender of the ploy of funny instantiation will have no trouble finding further supposedly absurd exampled of organizational teleological properties in non-biological systems.

I reject the ploy of funny instantiation nonetheless. It relies solely on unfounded intuitions, namely that this and that kind of system could not be conscious. The intuitions are unfounded insofar as there are no reasons to expect that human common sense, evolved in the context of middle-sized objects moving at speeds far from the speed of light, social groups of around a hundred people and so on, to be a good judge of what could and could not be conscious. Far from it, if we *had* the intuitive ability to decide whether an arbitrary physical system could be conscious or not, this would itself constitute an independent philosophical and scientific mystery. How did we obtain such a mysterious capacity?

Intuitions are forged in the fires of our ancestor's fight for survival and our everyday lives, and neither we nor (I suspect) our ancestors regularly encountered brains composed of a billion people, beer can computers and the like. This is why our naive intuitions at this point arguably are no more reliable than our naive intuitions about the shape of the earth or the size of the moon.

It is a commonly acknowledged problem of arguments mainly based on intuition is that it is hard to decide whether a given intuition is "genuine" common sense or whether it is an acquired taste, so to speak, something professional philosophers got used to in the turn of their careers. On this point I can claim to have some record of my own pre-philosophical intuitions about scenarios like Ned Block's Chinese brain. In my youth I have written a science fiction short story where a galaxy spanning civilisation develops a conscious super-mind that is constituted by its members playing the role approximating that of neurons. For what it is worth, the proposition that the Chinese brain could be conscious, to my younger self, didn't seem like a terrible violation of common sense but a rather straight-forward consequence of a broadly naturalist understanding of the mind.

6.7 Summary and Open Questions

We have introduced a metasemantic account of first-tier representational properties. Our approach has been motivated by Bayesian cognitive science, but also may be thought of as an application of *a priori* principles as a guide to the study of neuronal processing.[28] The upshot is that, to represent the causal structure of the environment, a system must consist of a network of internal states whose internal causal relations are globally isomorphic to the probabilistic relations of states in the environment. The structure of the isomorphism is dictated by Bayesian probability theory. Thus put provocatively, probability theory dictates constitutive laws of perception.

The resulting view is a form of generalization of Gentzen-style inferentialism to inductive inferences. Such an inferentialism holds that systems represent their environment as a probabilistic source of sensory input. The resulting view is a form of computationalism about content that squares well with the Bayesian approaches in cognitive science. The following chapter will apply these insights and develop a representationalist account of consciousness.

[28] Jaynes 1988.

7 Qualitative Consciousness

Part one argued that consciousness may be explained entirely in terms of representational content. While this ought to be good news for the naturalist, because representational content seems less inexplicable in terms of physical properties than phenomenal character, chapter five argued that it is precisely the peculiar nature of conscious content that threatens classical naturalistic accounts of the representational. We began to alleviate this flaw in chapter seven, where we saw how representational properties can be explained in terms of a state's role in unconscious approximate Bayesian inference. In this chapter we will investigate what kind of representational content Bayesian inferentialism predicts and explains. This will enable us to see that the content of Bayesian perceivers maps well onto the phenomenologically manifest content of perception. Thus, this chapter will begin to develop our take on the naturalization of consciousness.

The previous chapter argued that representational systems represent the causal structure of their environment in virtue of their internal computational structure. In particular, external states of affairs are represented as causes of current and expected sensory input. In this section we will investigate whether this picture succeeds in capturing the phenomenally manifest content of perception.

At first pass, phenomenologically speaking, there is something to the idea that perceptual states represent the causal structure of the environment causing sensory input. In trying to analyze the content of visual experience, Searle proposes the following:

> The Intentional content of [a] visual experience therefore has to be made explicit in the following form: "I have a visual experience (that there is a yellow station wagon there and that there is a yellow station wagon

> there is causing this visual experience)." This looks puzzling, but I think it is on the right track.[1]

Thus Searle holds that when we ordinarily experience an object, the fact that this object is the cause of our seeing it is baked right into the content of the experience. To see an object *is* to see it as causally interacting with one's own visual apparatus. If this analysis is phenomenologically plausible, which I would agree with Searle that it is, this speaks to the general phenomenological plausibility of the Bayesian view of perception.

Giving an account of how one represents the causal structure of one's environment is necessary to account for conscious perception, but it is not sufficient. When seeing a yellow station wagon I not only see it as an environmental cause. Rather, I see it as bearing appearance properties, in particular, the station wagon appears yellow. What can the Bayesian account of perception tell us about how our experiences come to represent appearance properties?

The central thesis of this chapter will be that we can capture the representation of appearance properties in terms of the inferential roles of sensory low-level representations in the predictive hierarchy. These states function as semantic primitives in the inferential processes of Bayesian representational systems. All predictions can ultimately be cashed out in terms of their relations to representations at the lowest level. These states themselves serve as the inferential bedrock of the whole process. Figure 7.1 serves as an illustration of the general idea.

So here is our first pass at *Bayesian representationalism*, the view that conscious content is explained by the Bayesian inferential hierarchy realized in the brain as captured by Bayesian inferentialism. The brain represents the probabilistic and causal structure of its environment insofar as this environment impinges on the organisms sensory surface. The sensory states themselves represent appearance properties of the causes in the environment. For instance, when seeing the blue sky, the sky is represented as the cause of a particular appearance of blueness. When seeing a yellow station wagon, the station wagon is represented as the cause of a particular appearance of yellowness.

[1] Searle 1983, p. 48.

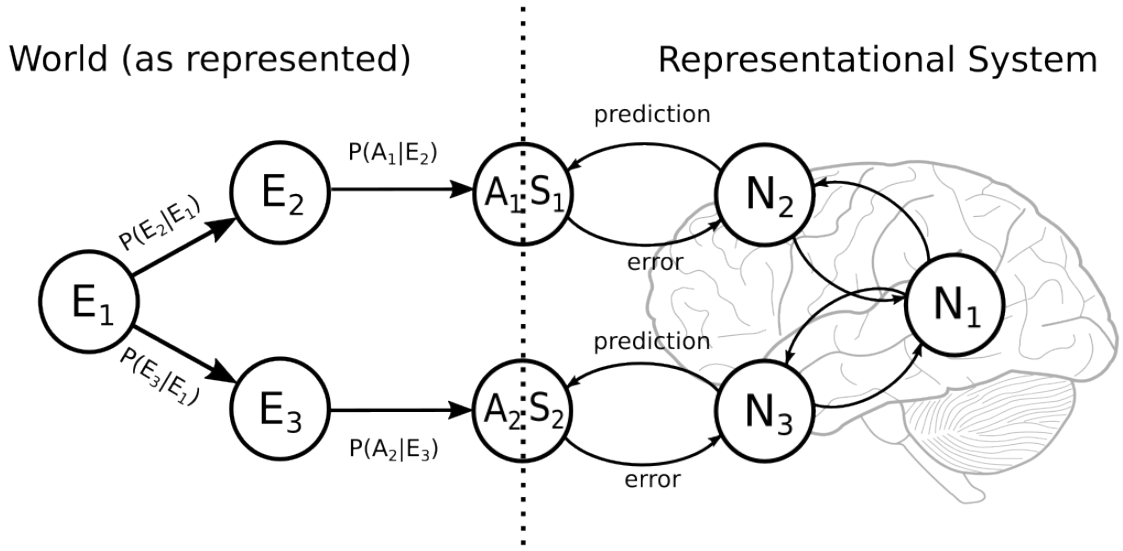

Figure 7.1: Schematic illustration of a representational system and the environment it represents. The dotted line marks the system-environment distinction. States N_1 to N_3 are neuronal states internal to the system that represent environmental states E_1 to E_3. Low-level sensory states S_1 and S_2 represent probability distributions over appearance properties A_1 and A_2. Arrows on the left ("in the world") depict probabilistic dependencies that may be expressed as conditional probabilities. Arrows on the right ("inside the agent") depict error and prediction signals. Note that unlike in figure 5.4 error and prediction units, which are typically thought to be realized by separate neurons, have been drawn into a single element. Also this depiction abstracts away from active inference which, according to our interventionist take on causal relations, would ensure the proper representation of the world in causal rather than mere probabilistic terms.

The investigation into consciousness in the chapter will be partial. In chapter two, following Kriegel, we differentiated between the qualitative and the reflexive nature of consciousness. Qualitative consciousness describes the awareness of an intentional object. Reflexive consciousness describes the fact that phenomenally conscious states represent their intentional object *to a subject*. We argued that this aspect of consciousness can be captured by a form of self-representationalism, the view that phenomenally conscious states represent, among other things, themselves.

In this chapter we will focus on explaining the qualitative aspect of phenomenal consciousness. Thus we will discuss how consciousness represents environmental states where environmental states may include the state of the organism itself, as in the case of pain experience. We will thus bracket the self-representational nature of conscious experience as well as the epistemology of consciousness. The coming chapter will then complete our account of consciousness by providing an account

of why consciousness is in some sense epistemologically transparent and why its states are inherently reflexive. For now, let us begin by investigating the relation of appearance properties and probabilistic representations.

This chapter will explicate Bayesian representationalism as an account of qualitative consciousness more detail. Section one will address the connection between appearance properties and low-level representations. Section two will address the phenomenological challenge of explaining why consciousness doesn't seem probabilistic.

7.1 Bridging the Phenomenology-Content Divide

This section will argue that the lowest states of the predictive hierarchy represent appearance properties of the ostensible causes of sensory input and address some issues regarding the phenomenological plausibility of this view. The discussion will be divided into two subsections. First, I will argue why the thesis that low-level predictions represent appearance properties can be justified with reference to the inferential role of these low-level representations. Secondly, I will discuss how Bayesian representationalism can address the complex structure and diversity of experience.

7.1.1 Representing Appearance Properties

How could we discover whether low-level representational states represent appearance properties? Inferentialism suggests that it should in principle be possible to read the representational import of some state off its inferential role. This is what we will attempt here. To do this we first have to investigate the inferential role of low-level representations in the predictive hierarchy.

Dialectically, my argument for the identity of appearance properties with the representational content of low-level representations is somewhat tricky. On the one hand, appearance properties are primitive and seem unanalyzable. It seems there is nothing we can do to elucidate or describe them: You have to experience them for yourself. Nonetheless, experiences of appearance properties play a definite role within our mental economy. May arguments will thus try to show that the cognitive role of experiences is precisely the inferential role of low-level representations.

In the static situation, meaning in systems that minimize their prediction error rather than their expected prediction error, low-level representations mediate between the external world and internal processing. The essential situation is depicted in figure 7.1. Representations at the lowest layer encode sensory input from the environment. The predictive hierarchy then performs approximate Bayesian inference by minimizing prediction error such as to find an approximate posterior, given the priors encoded at higher levels. Importantly, it is the lowest level in the hierarchy that drives all hypothesis revision.

Let's now investigate the inferential roles of low-level representations in the dynamic model. To repeat, in order to deal with the uncertain future it is imperative to minimize prediction error. This, on the face of it, is impossible because future prediction error depends on future sensory input which is unavailable in the present. So instead of minimizing future prediction error directly, we have shown that one can instead make a best guess about the future and minimize expected prediction error instead.

In a system that minimizes expected prediction error, low-level representations will play a double role. On the one hand, just as in the static case, they will play the role of sensory states that mediate between system and environment. But furthermore, low-level representations will represent *expected sensory states* that serve as the basis for calculating expected prediction error. This suggests that in the dynamic picture agents represent their environment not merely as the source of current sensory input but as the source of expected sensory input.

In both the static and the dynamic construal low-level representations serve as the ground for representational content of higher-level representations. Higher-level representations can be analyzed in terms of their predictive relations towards states at lower levels. This analysis however has to stop at the lowest level: low-level representations themselves serve as semantic primitives 'from the perspective of the system'. We have defined appearance properties as primitive ways of appearing or as primitive ways of being represented. It is thus not unreasonable to suspect that low-level representations represent appearance properties.

There are three ways of arguing for this convergence in a little more detail and I will now go through each of these. The first argument is phenomenological. When we take any object of experience and try to analyze it, the experience can typically

by decomposed into more basic constituents. As Peircean transparency suggests, these constituents will correspond to more basic properties of the intentional object of the experience. For instance, when attending to my coffee mug, the experience can be analyzed in terms of experiencing a certain shape, texture and color.

Now it seems that this kind of analysis will come to an end somewhere. The compound property of cupness is intuitively phenomenologically analyzable into elements, and maybe the elements can be further analyzed in the same way. The basic constituents however, the experience of color and maybe basic intuitions of spaciousness (though this spaciousness arguably cannot exist without content and is thus not an appearance property strictly speaking), cannot be further separated into more basic elements. These basic elements, we argued in chapter two, are best understood as appearance properties, primitive ways of appearing.[2] In this sense appearance properties are the unanalyzable phenomenological bedrock of experience.

We can arguably make sense of the phenomenological decomposition of objects of experience in terms of the predictive hierarchy. The perceived objects can be conceived as hypotheses represented at a mid-level of the predictive hierarchy that explain away activity at lower levels.[3] Remember that attention was explained in terms of expected precision which in turn can be conceptualized as a volume control on prediction error. Thus attending to the object entails higher prediction error which leads to a revision of the hypothesis that the input is caused by a single coherent object. When mid-level predictions get revised as prediction error increases lower-level predictions become salient. This decomposition of high-level predictions into underlying predictions under the scrutiny of increased expected precision plausibly parallels the decomposition of phenomenological objects into their elements.

Now just as the phenomenological decomposition comes to an end in basic appearance properties, the decomposition of hypothesis into more fundamental ones comes to an end at the bedrock of the inferential hierarchy. This suggests that the representational content of low-levels of the hierarchy should be identified with the basic phenomenological constituents, i.e. appearance properties. Of course the force

[2]Note that the idea of a phenomenological part is not quite clear. That is, the current analysis does not commit to any strong views on the relation of the cup and its phenomenologically more basic components. The claim here is just that there is some intuitive sense in which some experiences are analyzable into constituents.

[3]Hohwy attempts to explain binding, the phenomenological synthesis of many elements of experience into an object as a kind of inference to a common cause (Hohwy 2014, chapter 5).

of these phenomenological considerations is limited because they rest on a particular interpretation of the relation of the predictive hierarchy and the phenomenology of perception. Still, I presume it makes an intuitive case for the view that appearance properties are represented at the lowest levels of the predictive hierarchy.

The second argument for supposing that low-level representations represent appearance properties is semantical. It argues directly that the truth conditions of satisfaction of representations of appearance properties mirror the truth conditions satisfaction of low-level representations of the predictive hierarchy. As we saw before, the truth conditions of representations of appearance properties cannot be spelled out non-circularly. An object bears a particular appearance property if it in fact appears in a particular primitive way.

The truth-conditions of low-level representations are best captured in a similarly circular way. First, consider the static case. Here it is most reasonable to hold that low-level sensory states are in fact self-verifying in the sense if a sensory state is in x then it is veridical precisely if it is in x. Prediction error generated at the lowest level is not the result of sensory states themselves but of the mismatch between predictions coming in from higher levels and the relevant sensory states: Sensory states themselves can never be false in any relevant sense.

In the dynamic case the situation is more complicated but similar in essence. Low-level representations here are not self-verifying because expected sensory states can of course fail to come to pass. Still, the truth-conditions of these states are best spelled out in a circular fashion. If, for instance, the some low-level representation is predicted to be in state x, then the particular prediction will be veridical precisely if the future sensory state is x, non-veridical if not. No further spelling out of the truth-conditions is possible.

All this is to say that, just as appearance properties are ways of appearing for which no non-circular analysis can be given, low-level representations of predictive processing systems are best construed as semantic primitives that also cannot non-circularly analyzed semantically. Low-level representations function as semantic primitives that attribute primitive properties to the causes of sensory input.

The third argument for the view that low-level representations represent appearance properties appeals to the special epistemological status of appearance

properties and low-level representations. Arguably, knowing that a certain object has a certain appearance properties entails minimal knowledge about that object. We called this characteristic the phenomenological thinness of perceptual experience. Knowing that something is appearance-red merely entails that it is currently represented in *this* particular fashion.

Low-level representations in the predictive hierarchy mirror this epistemological peculiarity. Because all 'knowledge' about the world will be encoded in higher-level priors and predictions, the epistemological import of low-level representational states will be minimal. Compare: Activity at higher level in the hierarchy generally involves complex commitments about sensory input. The same cannot be said about low-level representations.

A referentialist may want to hold that low-level representations represent direct sensory stimulation in some fashion and that sensible properties are physical properties of some kind. From the (internalist) inferentialist perspective this view is suspect because low-level sensory representations certainly do not *transparently* represent physical properties. Transparency is here used in an epistemological sense where a mental representation represents some content transparently if it is clear to the subject of the representation that this is in fact the content. After all, that sensible properties correspond to specific physical causes in the environment itself is something that has has to be inferred. The inferential role of low-level sensory states thus cannot be such as to represent anything but primitive properties that serve as the ground of inference.

A forerunner of the view that the peculiarities of low-level representations in perceptual inference may explain how perceptual states come to represent primitive sensible properties. In his *Imaginary Foundations* paper he has suggested that systems processing data in a Bayes-optimal fashion actually need to represent their own sensory states with probability that is equal to unity. On his account these sensory states are best thought of as representing what he calls 'imaginary propositions', propositions that serve a special role in inference but that do not pertain to objective matters of fact.[4] The current account is greatly inspired by Schwarz's proposal but it is not essentially committed to the assumption that perception is best framed as ideal Bayesian inference. Also, according to Bayesian representationalism,

[4]Schwarz 2018.

while the content of low-level representational states pertain to what we may call relational properties they are still perfectly objective.[56]

Finally, the inferential role of low-level representations explains why inferentialism avoids the mismatch problem faced by referentialist metasemantics, particularly structuralism. The problem of structuralism was that appearance properties can't be characterized in terms of their relational properties alone. The inferentialist is not committed to representations which's contents can be analyzed purely in terms of relational properties. Consider, for instance, indexical terms. Indexical terms clearly can't be analyzed purely in terms of the relations of the objects they are applied to. Still it is plausible that the meaning of indexical terms can be analyzed in terms of their inferential roles. In this way, the semantic primitives represented by low-level representational states in the inferential hierarchy can be likened to indexicals.

An important worry about the view that low-level representations represent appearance properties remains. I have argued in chapter two that the representation of appearance properties implies a form of self-representationalism. So far however I have not addressed how the inferential roles of low-level representational states explains the self-representational nature of appearance properties. I acknowledge that this is an important problem, but one best bracketed for now. In the coming chapter we will discuss in some detail how predictive processing theories can account for reflexive nature of experiential states.

The suggestion that low-level representations in hierarchy inference represent appearance properties is promising. If correct we would have made important leeway in understanding consciousness from a naturalist representationalist perspective. Furthermore we would have demonstrated that inferentialism has a definite advantage over its referentialist rival as, whereas the latter cannot account for appearance properties at all, the former has a natural place for them. In the following section we will flesh out emerging account in a little more detail and see whether it has an illuminating story to tell about the complexity of conscious phenomenology.

[5] ibid. conceived of imaginary properties as non-objective in the sense that they do not bear logical relations to other proposition to other propositions. But this is incoherent. Every imaginary proposition at least logically implies that an agent represents something in some primitive way. Thus, Schwarz's proposal seems to collapse into the appearance account.

[6] Another kind of forerunner of the ideas of appearance properties as semantic primitive in an inferential hierarchy are theories that tried to account for conscious sensations as semantically primitive lexemes in a language of thought approach (Rey 1991; Leeds 1993; Lycan 1996).

7.1.2 The Intrinsic Structure of Experience

Conscious phenomenology has a complex internal structure. First of all, experience acquaints us with a huge variety of different appearance properties that themselves have a rich internal structure. It seems to be an intrinsic feature of color experiences to stand in relations of similarity and difference towards each other. A red experience seems intrinsically more similar to an experience of orange than to an experience of blue. Every explanation of conscious phenomenology must involve an explanation of the intuited similarity between the colors. Thus Bayesian representationalism must give some account of how this internal structure comes to be as a result of approximate Bayesian inference.

Secondly, conscious phenomenology also seems intrinsically perspectival. We experience the world, as it were, from a special vantage point, the point of view of the observer. While it may sometimes be hard to locate this vantage point exactly it normally is somewhere inside the head. All contents of experience are typically experienced relative to this indefinite point of origin. Even when imagining something the imagining will typically share the perspectival structure of perceptual experience. Ideally, Bayesian representationalism should have something to tell about the perspectival structure of experience.

Thirdly, in chapter three we saw that consciousness has a field-like structure that accounts for our experienced unity of consciousness. There we argued that the sense in which a number of synchronous experiences are typically experienced as part of a single unity is best spelled out as a particular kind of representational content. In particular, the relevant intentional objects are experienced as part of a single space. We connected this experienced space to the intuitive sense in which experiences form a unified phenomenal field. Bayesian representationalism should have something to say about this spatial structure of experience.

Finally, in the same chapter we argued that consciousness is typically experienced as a kind of stream. This stream of consciousness is best captured by the fact that the content of consciousness is typically temporally structured. This represented temporal structure, we argued, also partly accounts for the sense of unity of experience. Even two disparate experiences like a taste of coffee and a pain in the knee can be

experienced in temporal relation towards each other. Bayesian representationalism should have something to say about the temporal structure of experience.

Generally Bayesian representationalism can account for these complexities of experience in terms of the structure of the generative model (denoted by M in chapter five). The generative model constitutes the background assumptions about the world encoded in prior probabilities of a representational system. By selecting a suitable model we can make sense of the phenomenological structure of experience in terms of implicit assumptions about the world encoded in the model.

First, we can capture the fact that there seems to be a fixed number of appearance colors by presupposing that there is a similarly fixed number of types of low-level representational states as specified by the generative model. It then becomes straightforward to address the fact that experienced colors have an intrinsic structure of similarity and difference. We can capture this structure by assuming that the relevant low-level sensory representations are associated with priors such as to make unconscious inferences to similarity and difference more or less likely. For instance, the fact that a red experience is intuitively more similar to a orange experience than to a blue experience may be explained by the fact that visual processing is disposed to attribute red and blue-stimuli to different environmental causes while it is disposed to attribute red and orange ones to the same cause. In this way we can make sense of the structure of color experience in terms of the structure of the underlying generative model.

Note that the prediction error formalism is neutral on the question of which parameters of the generative model can be modified by learning. The formalism merely tells one which parameters one has to tweak to minimize expected prediction error. In systems that are capable of learning, the formalism can capture this fact by holding that the prior-encoding parameters are changed such as to minimize expected prediction error.[7] But if some property of the model is not subject to learning then we can assume that it is not subject to such change. Thus the suggested account can deal with the apparent constancy of aspects of experience like intuitive similarity relations between colors.

Of course this is in some ways an *ad hoc* response to the problem of the phenomenological structure of color experience. As such, it lacks experimental backing

[7] Friston 2010.

and integration with what we know about the neurophysiology of color vision. But the point here was not to offer a full model of the nature of color experience but to demonstrate that Bayesian representationalism has the required conceptual resources to make sense of experiential structure in representational terms. In particular, the intrinsic structure of experience can be handled in terms of the shape of the system's generative model. Not only does the model define what possible sensory states there are, but it also determines the intrinsic structure of the properties represented by the sensory states. So the structure of color-space is just one example of the structure of appearance properties. The account can obviously be generalized to cover other sensory modalities.

Appearance properties have a strange dual role. On the one hand they are experienced as attributes of the environment. On the other hand they relate the environment to the structure of the agents representational system. Above we tentatively expressed this by noting that appearance properties may be called relational properties of the objects of experience - relational with regards to the experiencing agent. It thus should not come as a surprise that the characteristics of appearance properties, like their intrinsic structure of similarity and difference, is determined by properties agent's generative model. If appearance properties are primitive ways of being represented then the character of the representational system in question will partly determine the character of appearance properties.

I now want to discuss a more complex and scientifically advanced model of the intrinsic structure of human experience. The *projective consciousness model* tries to explain the spatial and perspectival structure of experience in terms of the integration of ideas from Bayesian cognitive science with the proposal that the brain utilizes the structure of projective geometry to organize information. If correct, this model offers a powerful account of the structure of experience. Furthermore, if it is correct that phenomenal unity can be explained in terms of represented space, the model also has the potential to address the phenomenological datum of phenomenal unity.

At the heart of the projective consciousness model lies the insight that the geometry of experience is not actually ordinary euclidean geometry. In euclidean geometry there are no privileged points of origin. On the other hand, the geometry of experience is in an important sense subject-centered. Objects inherently appear from a particular

perspective. We never experience front and backside of an object in quite the same way. Furthermore, objects appear larger when closer to the observer. All these features are not well captured in a euclidean model of experienced space.[8]

Projective geometry studies the geometry spanned by a set of lines through a privileged point of origin as illustrated in the depicted image.[9] This kind of geometry can thus be seen essentially as the formalization of the Renaissance concept of a vanishing perspective and one may argue that the artistic appeal follows directly from its accurate depiction of the structure of experienced space.

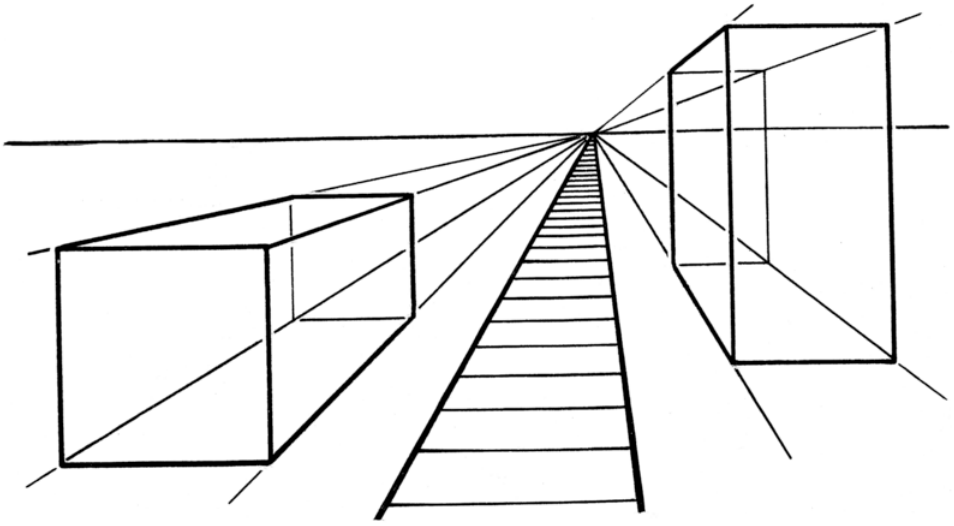

Figure 7.2: An illustration of the vanishing point perspective of experience.
An illustration of the vanishing point perspective of experience as an example of projective geometry. The two-dimensional perspectival projections of parallel lines appear to converge.

The projective conciousness model is an ambitious integrative model of consciousness. It tries to account not only for the spatial phenomenology of experience but also for the phenomenon of imagination, imagined perspective taking and experienced valence. But at its heart lies the simple insight that the geometry of experience is best described by projective geometry underling neuronal representation.

According to the projective consciousness account our brain's generative model is best thought of as consisting of two components, a non-perspectival model S and a variable projection operator T. The generative model M will then be separable as ST. A prediction error minimization will then result in selecting a particular projection

[8]Rudrauf et al. 2017; Williford et al. 2018.
[9]This image is available under a creative commons licence on commons.wikimedia.org.

such as to achieve an optimal organization of the non-perspectival information encoded in S.[10] What does this mean without mathematical jargon?

Imagine sitting at your desk engaged with some task. Your focus, we assume, lies on your laptop screen and you experience space from the vantage point of your skull. Suddenly you hear a sound behind you even though you think you are alone in the room. Before your eyes flash the image of the door behind you opening and someone entering the room. In this imaginative episode, you experience the situation from another perspective. Maybe you experience it as though you already turned your head, or from some disembodies perspective in the middle of the room. You may then turn your head to see whether it is really the case that someone has entered.

In the projective consciousness model this process of perspective taking is best described as a minimization of prediction error over a number of possible perspectives, characterized by projection operations. When you are focused on what you are doing with your hands the perspective, i.e. the projection operator that minimizes prediction error will correspond to a perspective from behind your eyes. When you hear a sound behind you the perspectival character of your experience changes because the information that is currently most relevant to the minimization of prediction error will now be information about what is happening behind your back.[11] The projective consciousness model thus gives an elegant interpretation of the phenomenology of perspective taking in terms of the minimization of prediction error.

Importantly, projective geometry enters into the account merely in terms of a mathematical operation. This operation can ultimately be cashed out in terms of the connection strengths inside the hierarchical model of the brain in a similar way as in chapter five we saw the linear structure of time may be so encoded. Thus the model is potentially a powerful tool for accounting both for the perspectival and the spatial structure of experience. If the model is correct, both result from the fundamental structure of the brain's generative model and it's separability into non-perspectival information and a variable projection operator.

Defenders of the projective consciousness model hypothesise, as the title already suggests, that there may be a constitutive connection between the availability of a content

[10] Rudrauf et al. 2017.
[11] Ibid.

to conscious introspection and its representation under a specific spatial projection.[12] The authors further speculate that projective geometries may be connected to the global neuronal workspace, a topic we will discuss below. At any rate, it seems to me that these speculations, based on phenomenological observations, are insufficiently justified. The authors seem to be subject to a simple phenomenological fallacy. While it seems to be the case that all phenomenologically conscious content has a perspectival structure this does not entail that the same isn't true for unconscious contents as well. Thus, the perspectival structure does not explain phenomenal consciousness.

Defenders of the projective consciousness model have committed Julian Jaynes' *flashlight fallacy*: The fallacy of conflating introspectable features of the conscious mind with its essential features. This is a fallacy because unconscious processes are intrinsically non-introspectable. Concluding on phenomenological grounds that perspectival representation is characteristic only of *conscious* representation is like wandering through your dark attic with a flashlight and concluding that its bright everywhere because you pointed your light wherever you looked.[13] I will present a more principled model of the conscious-unconscious divide below.

Another issue is whether one should interpret the projective consciousness model in a thoroughly representationalist manner or whether the spatial nature of mental representation ought not partially be spelled out in terms of a direct awareness of the vehicular properties of mental representation, i.e. as a form of mental paint. Some defenders of the model have recently favoured an anti-representationalist understanding on the grounds that to suppose otherwise would entail that conscious mental representation is largely and systematically falsidical. After all, the space we inhabit is (putting Einsteinian considerations aside) Euclidean. If it is part of the content of visual experience that the geometry of space is projective rather than Euclidean then the relevant content must be falsidical. If one holds that such systematic illusions are implausible then it may seem desirable if one could interpret experienced space in some anti-representationalist fashion.[14]

It is natural at this point to suppose that the mentioned problem can be solved by appealing to appearance properties. I argued that experienced color properties are not

[12] Rudrauf et al. 2017; Williford et al. 2018.
[13] Jaynes 1976.
[14] Williford et al. 2018 p. 12.

mere physical properties that pertain to the object of experience independently of the experiencing subject. I also suggested that the intrinsic structure of color-space is best explained by the properties of the agent's generative model. In the same way, we can capture the nature of experienced space in a representationalist way by supposing that spatial perspective objects are experienced in terms of are best spelled out in terms of appearance properties that are structured in part by the agent's generative model.[15] On this account, we experience the world *as appearing* in a certain spatial manner.

The projective consciousness model is speculative in nature and thus I will not base my following account on its validity. Rather, the purpose of the previous presentation was to illustrate that Bayesian representationalism offers a powerful framework for thinking about the nature of consciousness while leaving many details to consciousness science to figure out. The projective consciousness model offers a good example of how future cognitive science may try to fill in the details.

As a final point let me briefly expound on how Bayesian representationalism can account for the phenomenology of time. In light of what we already said, this explanation is straightforward. In chapter five we saw that the free energy principle can deal with temporal depth by assuming a model that intrinsically runs into the future to some predefined time-horizon. The formalism of expected free energy then results as a natural correlate. All this implies that the phenomenology of time can be explained in an analogous way as we explained the phenomenology of color and of space: We just have to assume that temporality is build into the structure of the generative model.

At this point it should be noted that the shape of the generative model our brains use to capture time is a debated issue. In chapter five we discussed a potential discrete model of temporality. Probably continuous models where neurons code for rates of change of environmental variables are more realistic.[16] The latter approach may also explain why phenomenological time seems continuous rather than analyzable into some smallest interval.

The view that the structure of experience is to be explained in terms of the shape of an generative model presupposed in approximate Bayesian inference is as much an at-

[15] ibid. partly build their critique on Thompson 2008 who argues that an appropriate theory of consciousness needs to involve mental paint. However, he accepts that one way of escaping this view is to hold that the properties of the intentional objects of experience are not wholly subject-independent (ibid., p. 394). This is precisely what I am arguing.

[16] For discussion, see Parr, Pezzulo, and Friston 2022, chapter 4.

tempt at providing a philosophical theory as it is a framework for scientific research. It would thus be presumptuous to attempt to provide a more detailed account of what exactly such an generative model should look like. Should Bayesian representationalism turn out to hold any merit then giving such an account will be a task for consciousness researchers of the future. The projective consciousness model, if valid, will turn out to be a huge step in that direction. For now let me bring the discussion to a close and address an important phenomenological objection to Bayesian representationalism.

7.2 "...why doesn't it *seem* probabilistic?"

A recent paper by Ned Block is entitled *If perception is probabilistic, why does it not seem probabilistic?* Block hints at the fact that on typical Bayesian accounts of perception states of the world are not just represented as being in some specific state, but they are represented to be in a variety of states with certain probabilities. But in conscious perception, the states of the world appear determinate.[17] I see my cup standing on the desk in a determinate manner. But if Bayesian representationalism is correct, the relevant perceptual state should really represent the cup as being in various states with various different probabilities. How is the apparent mismatch to be explained? Why don't I see the cup as a probabilistic blur?

There are two general strategies to divert Block's criticism. First, one may argue that perception actually seems more probabilistic than Block assumes. Secondly, one may argue that where consciousness does not seem probabilistic, this can actually be explained in terms of Bayesian representationalism. Both replies have an element of the truth. I will discuss them in order.

I can think of at least two ways in which it is plausible to argue that consciousness actually *does* seem probabilistic. One is consciousness of potentialities. As an example, take the often made remark that three-dimensional objects we see typically *look like* they have a back side, even if we do not currently see it. How can we explicate this vague phenomenological intuition? It seems that the intuition is partly explained by an expectation regarding what we would see, given we were able to see the object from the other side. Now it seems also plausible that such consciousness of

[17] Block 2018.

potentialities, consciousness of what *would be* given some counterfactual perceptual scenario, is probabilistic in nature. When I look at my cup, I am aware of the fact that if it were turned around, I would be able to see its back. But being so aware I am not aware of any specifics. For instance, it is not that my vague sense of what I would see there includes the coffee stain on the back side of the cup. But it is also not the case that I represent the cup as *not* having the coffee stain. It seems that consciousness of potentialities is inherently indeterminate. These indeterminacies of the content of experience can be spelled out probabilistically. When I turn the cup and see the other side, different ways the cup may look can be rank-ordered in terms of their degree of conformity with the original impression of the cup as having a back side. The other side with and without the coffee stain will be in approximately equal conformity with my original impression. If the back-side turns out to be made of glass rather than ceramic, this would partly invalidate the original sense of a back side. If the cup turns out to be a two-dimensional dummy, the sense of it having a proper back side similar to its front was utterly illusory. Thus there seems to be a rank-ordering of different states of the world that are more or less in accordance with the visual impression. Such a rank-ordering can arguably be spelled out as a probability distribution over various possibilities that constitute generalized truth-conditions of the perceptual state. Thus the perceptual state of the cup that includes a vague sense of a back side will be more or less accurate (rather than just true or false *simpliciter*) depending on the cup's back side.

We can also argue for the indeterminacy of the consciousness, and thus its probabilistic structure, in a more direct fashion. An intimation of the probabilistic structure of experience may be gleamed by considering the phenomenology of vision in the corner of the eye. Eric Schwitzgebel has argued that here introspectable phenomenal facts are seemingly underdetermined in the sense that the actual phenomenal character is actually hard to pin down introspectively.[18] From an representationalist vantage point, this is best explained as a case where the representational content is not fully determinate. On the other hand, it does seem that we can perform probabilistic judgements about the things we see in the corner of our eyes. When I see book lying there in the corner of the eye, this impression gives me some intimation of how large

[18]Schwitzgebel 2006.

it may be. It is quite plausibly no larger than such and such, and almost certainly smaller than such and such. As probabilities, as Cox taught us, are just numeric measures of plausibilities, these judgements indicate that the representation of the book is indeed best cashed out probabilistically. The corner of the eye is an excellent example of the presence of probabilistic elements in the content of perceptual experience.

We have thus seen that it is not entirely accurate to say that consciousness does not seem probabilistic. But still, the fact that at least at the center of attention consciousness is typically quite determinate remains unexplained. How may Bayesian representationalism account for the fact that we are normally focally aware of a singular determine state of affairs?

There are a couple of points to be made here. First, note that we are dealing with approximate rather than exact Bayesian inference. In exact Bayesian inference every epistemic possibility will typically be represented with some non-zero probability. But remember that is precisely this excess of possibilities that makes exact Bayesian inference computationally intractable. The solution was to use an approximate posterior $Q(e)$ that is zero for a huge chunk of the space of epistemic possibilities. In effect, approximate Bayesian inference differs from exact Bayesian inference by a huge narrowing of the field of possibilities relevant at every point. As we hold that the approximate posterior encodes the content of consciousness Bayesian representationalism does not predict that all possibilities enter into conscious content. The predicted content thus is 'less probabilistic' than one may suppose from an orthodox Bayesian perspective.

Secondly, the precision-weighing model of attention nicely explains the fact that objects at the center of attention appear highly determinate. After all, attention is supposed to be nothing other than the effect of a second-order model that captures the reliability of first-order contents. Thus it should necessarily be the case that what lies at the center of attention appears more determinate, more certain, than what lies at the periphery. If we accept that perceptual consciousness at the periphery of attention is adequately captured in probabilistic terms, then the phenomenology of attention is adequately captured in terms of the precision weighing model. The center appears determinate, the periphery fades out into probabilistic indeterminacy.

Thirdly, an even more powerful model for the explanation of the determinate contents of experience has been suggested by Jakob Hohwy. According to Hohwy, the unified and determinate contents of experience should be explained in terms of active inference. A pure perceiver, Hohwy argues, may indeed perceive the world as a mere probabilistic blur. However, active inference necessitates settling for one *winning hypothesis* based on what future action policy one decides on. On Hohwy's model, the explanation of why consciousness, at least at the center of attention, does not seem probabilistic is straight-forward: Ordinary conscious content is always already the result of settling for a winning hypothesis in guidance of action.[19] I will elaborate this approach in more detail in the following chapter. This brings our discussion of the phenomenological adequacy of Bayesian representationalism with regard to qualitative consciousness.

7.3 Summary

In this chapter we introduced Bayesian representationalism, the view that the contents of consciousness are the predictions of a brain performing approximate Bayesian inference. As we saw, there are good reasons to believe that this view can serve as a phenomenologically plausible model of the conscious mind. In particular it helps us to make sense of the fact that consciousness represents the world in terms of primitive appearance properties, which we argued are encoded by low-level representations in the predictive hierarchy.

Furthermore we argued that Bayesian representationalism is quite flexible with regards to the exact structure conscious experience may turn to have. This is because the view is neutral with regards to the shape of the generative model underlying experience. Thus Bayesian representationalism is more like a framework for constructing future theories of conscious experience than such a theory itself. As an example of what a scientific filling in of details may look like we discussed the projective consciousness model as a promising example of an account of the spatial nature of experience.

So far a number of important questions have been left unaddressed. In particular, we have not discussed the epistemology and the self-representational nature of

[19] Hohwy 2014, chapter 9 and 10.

consciousness. We will do so in the coming chapter. With the tools developed in the process we will also be able to more fully appreciate the difference between concious and unconscious representational processes.

8 Reflexive Consciousness

On first approximation it seems that, while knowledge of the external world is causally and inferentially mediated, knowledge of our own mental state is unmediated and direct. No effort is involved in finding out what one is thinking, seeing, believing, intending and feeling. All this is especially true for conscious experiences. It is plausible to hold that it is of the nature of consciousness to be knowable in this direct way. In this chapter will investigate the nature of phenomenal self-knowledge and its relation to the nature of consciousness. Only in this way it is possible to get a deeper understanding of phenomenal consciousness.

Self-knowledge, the knowledge of one's own mental states, is characterized by a number of interesting features. First, self-knowledge is *privileged* in the sense that one's own judgements are typically more probable to yield knowledge than are the judgements of other agents. Secondly, self-knowledge is *peculiar*.[1] It seems, on the face of it, wholly unlike our normal ways of attaining knowledge of contingent matters of fact. Self-knowledge usually does not seem inferential in the sense that we derive it from other facts. Also, it does not seem perceptual in the sense that we perceive our mental states in the same sense we perceive the external world. If the so-called *inner-sense theory* of self-knowledge, the theory that holds that self-knowledge can be likened to perceptual knowledge, should turn our to be correct, then our inner sense must be quite unlike other sense modalities. This is further evinced by the fact that introspection typically is not accompanied by special introspective phenomenology comparable to visual, auditory or olfactory phenomenology. On the face of it, introspection does not appear like just another sense modality.[2]

Furthermore, self-knowledge is typically held to be very *reliable*. We do not often feel forced to revise our judgements about what we are thinking about or

[1] I here use the terminology of Byrne 2001.
[2] Shoemaker 1994b.

what we are experiencing. Though I hasten to add that there are good reasons to hold that the impression of reliability of self-knowledge are often misleading. There are a variety of mental states that seem to be easily introspectable but our judgements about them are in fact quite fallible. Schwitzgebel has given the poignant example that his wife sometimes is better at judging whether he his angry than himself.[3] Also, psychologists have shown that introspectable intentions frequently may turn out to be *post-hoc* constructions.[4] And introspecting dispositional mental states like beliefs and desires is a notoriously deceptive business. To use a socially relevant example, one may sincerely profess not to hold any biases against some ethnicity while the actual judgements about members of these ethnicities clearly reveal such biases. As a general rule of thumb it seems that occurrent mental states are more easily introspectable than dispositional ones.

In our discussion we will focus on our epistemic access to phenomenally conscious states and touch on other mental states only in passing. All three features of self-knowledge apply directly to our knowledge of our own phenomenal states. The mode of access is privileged in that I can know the features of my consciousness much better than can third parties. The mode of access is peculiar, i.e. wholly unlike any other epistemic capacity I have in that it seems unmediated and direct. And thirdly, it seems to be highly reliable. In fact many would hold that knowledge of phenomenal mental states is in some sense infallible! Naive intuition would have it that when I contemplate a patch of green at the center of my visual field I could not possibly be wrong about what this experience is like.[5]

Schwitzgebel has pointed out that philosophers tend to overemphasize the degree to which we know our conscious experience. In fact, upon reflection, the precise phenomenal character of the periphery of our field of vision or the precise phenomenal character of emotional states is elusive. How exactly, for instance, does melancholy feel? Is it localized at some point in the body or is it not? But even Schwitzgebel would not deny that the phenomenal character associated with sensible properties at the fovea of vision is quite hard to get wrong.[6] A theory of the knowledge of phenomenal

[3]Schwitzgebel 2006.
[4]Nisbett and Wilson 1977.
[5]Interesting contemporary defences of this view can be found in Horgan, Tienson, and George 2006 and Horgan and Kriegel 2007.
[6]Schwitzgebel 2006.

mental states should have something to say about this gradation of introspective access. That is, it should explain why appearance properties seem easily introspectable while other features of phenomenal mental states seem to slip from grasp easily.

Our investigation into the nature of phenomenal self-knowledge is more than just an excursion unconnected to the project of Bayesian representationalism. Remember our distinction of qualitative consciousness, i.e. the property of consciousness to be directed at an object, and reflexive consciousness, i.e. the property of consciousness to exist for a subject. We have seen that qualitative consciousness can be explained in terms of approximate Bayesian inference regarding environmental causes of sensory input. We also hinted that reflexive consciousness can be explained in terms of the self-representational character of conscious states, i.e. their property of representing themselves. So far however we haven't located these self-representational capacities in our Bayesian paradigm. To offer a full understanding of consciousness we thus need an account of its epistemology and self-representational nature.

Further, in the previous chapter I argued that the fact that consciousness represents appearance properties can be explained by low-level representations in the inferential hierarchy. These, I argued, represent certain semantic primitives in virtue of their fundamental inferential role. However this does not suffice to account for the fact that they represent *ways of appearing*. Insofar as appearance is itself a semantic notion this would involve that the relevant semantic primitives are intrinsically represented *as semantic*. The representations in question must be reflexive. Accounting for the reflexive character of mental representations will be invaluable for explaining how neuronal states represent appearance properties in virtue of their inferential role.

The chapter will be structured as follows. First, we will introduce Jakob Hohwy's model of self-knowledge in the predictive mind. Section two will then show how we can employ this model to explain self-representational states and the reflexive nature of consciousness. However our model will entail that reflexive and qualitative consciousness are actually metaphysically independent inferential features of mental states and thus they should be able to exist independently of each other. Thus section three will investigate whether such dissociation of reflexive and qualitative consciousness is plausible. Section four will introduce the so-called winning hypothesis model of consciousness that will help us integrate our findings with contemporary conscious-

ness science and also draw a principled distinction between conscious and unconscious processing. Section five will reflect on the consequence of the emerging view that consciousness is not actually a natural kind but an amalgamation of a number of computational features of certain representations. In section six I will note some observations about unconscious mental processes. Section seven will discuss whether or not our model has anything interesting to say on the hard problem of consciousness.

8.1 Double Bookkeeping

In his seminal *The Predictive Mind* Hohwy proposes a model of how to think about self-knowledge in the context of predictive processing. My account will be based largely on Hohwy's. The general idea is that a prediction error at some stage in the predictive hierarchy is the result of two factors: Of the environment acting on the sensory surface *and* of the representational processes at lower levels of the hierarchy. Thus there are two ways of explaining away prediction error. On the one hand one may change one's views about the environment. This accounts for perceptual inference. On the other hand one may change one's views about the relevant lower-level representations. This second possibility accounts for introspective inference.[7]

Let us investigate two examples of how introspective inference may come about in the predictive hierarchy. First, imagine standing up after sitting at the breakfast table and feeling dizzy. You *could* interpret the unexpected arising spinning as an actual rotation of your body. This would certainly be one way of explaining away the arising prediction error. But in the light of your prior beliefs (one does not typically start spinning rapidly for no apparent reason) it is much more economical to suppose that the low-level inferential processes that are responsible for estimating your orientation are subject to error. Thus the relevant prediction error is better explained by an hypothesis about representational-inferential processes themselves.

To urge another example Hohwy uses to illustrate his point, it is a frequent symptom of an outbreak of psychosis that light appears too bright. Here too it seems that relevant prediction error could be explained away in terms of environmental causes: The light actually is brighter than it was yesterday. But of course, as we expect the

[7]Hohwy 2014 p. 245-249.

sun to be equally bright as yesterday, explaining the arising prediction error as a result of faulty internal to low-level perceptual inference is much more economical.[8]

What these examples suggest is that the general structure of self-knowledge is the following. Prediction error arising at some level of the hierarchy X can be explained away in two different ways. Either it is explained away in terms of an environmental cause or in terms of an internal mental or representational cause at some level beneath X. In the former case the higher level of the hierarchy comes to represent a worldly cause. In the latter case it comes to represent lower-level representations.

It thus seems that introspection can be explained quite straightforwardly in the predictive processing paradigm. Whereas perception is approximate Bayesian inference on environmental causes introspection is approximate Bayesian inference on mental causes. The predictive hierarchy will thus turns out to reflexively model aspects of itself. Following Lukas Schwengerer[9] we can call this the *double bookkeeping model* of self-knowledge.

There is an important difference between the double bookkeeping account and accounts of self-knowledge that are commonly discussed in the contemporary philosophy of mind. In the latter, the outcome of a process of introspection is typically a propositionally structured believe about one's own mental states. In the case of a predictive processing inspired view on introspection, introspection will result in just another addition to one's model of the sources of sensory stimulation, just that these sources of sensory stimulation will now the agent and her mental states. While I believe there is important work to be done on the mapping between propositionally structured mental representations and probabilistic models I will bracket the issue here. I will continue to treat self-knowledge as an aspect of a model of the sources of sensory stimulation.

It is important to note that the mode of epistemic access proposed here differs from the mode of access one has to environmental causes. The mental processes one introspects are not just further causes in the environment in that they are available in a more direct fashion. This explains why introspective inference is privileged. Only I (or 'my' predictive hierarchy) can infer that my low-level inferential processes have gone awry from the fact that the world appears to be spinning.

[8] Hohwy 2014 p. 246.
[9] Schwengerer 2019.

It also explains why introspective inference is peculiar because, as we have just explained, it differs from normal access to the world.

The double bookkeeping has to make sense of the reliability of self-knowledge in terms of the low ambiguity involved in the relevant inferences. Such low ambiguity is certainly plausible where our everyday repertoire of mental states is concerned. Here it is natural to hold that inference will appear to be subjectively certain. But if introspective inference is in some sense inductive one would expect that, in cases where novel mental states arise, introspection becomes a lot harder. But while this does not square well with the intuition that introspection is in some sense epistemically direct, it does seem to conform with our phenomenology: Introspection certainly is harder in cases where unusual mental states are concerned. In surprising and uncommon experiences it may even appear correct to say "I don't even really know how I feel right now." On the model we are considering this is arguably more than a mere figure of speech. It is a local breakdown of introspective inference in processing novel data.

The fact that self-knowledge does not *seem* inferential should not worry us too much. After all, as direct realists like to point out, perception also does not seem inferential either.[10] I look at the blue rose and see it directly. But this phenomenological directness is, Bayesian cognitive scientists would argue, mediated by unconscious Bayesian inference. If the double bookkeeping model is correct the same may be said about introspective inference.

Schwengerer has suggested that we can frame the double bookkeeping model as a variety of the *transparency account of self-knowledge*.[11] This account holds that we need not postulate a capacity of inner sense to explain introspection. We can explain it as a mere application of abilities we typically grant of rational beings. For instance, when applied to conscious experience, the fact that I see a cat may be directly inferred from the visual features of the cat. In a similar manner the fact one is dizzy is inferred from the fact that the world appears to be spinning. Mental states can be inferred directly from facts about the world because worldly facts are only available mediated by mental representations.

It is necessary to differentiate this claim of what we may call *introspective transparency*, transparency conceived as a model of how we usually attain self-knowledge,

[10] Huemer 2001.
[11] Schwengerer 2019.

from *phenomenological transparency*, transparency conceived as a thesis about the phenomenological structure of experience. Phenomenological transparency can be further differentiated into Moorean and Peircean transparency. Now while there is an intrinsic connection between introspective and phenomenological transparency claims they are certainly to be held apart conceptually.[12] The claim at issue here is whether the double bookkeeping model renders self-knowledge introspectively transparent.

While there is something to this transparency claim, this is not the complete story. Typically the inferences relevant to transparency will be deductive in the sense that the fact that one represents that P is inferred from the fact that P. While this is a strange kind of inference[13] it seems to be best captured as a kind of deduction. But the kind of inferential self-knowledge described by the double bookkeeping model is, contra Schwengerer, best described as inductive. In fact, there is no strong qualitative distinction between self-knowledge and perceptual knowledge in the model expect for the directness of access to its object and the level of ambiguity involved. Both are inferences to the best explanation. In some sense the double bookkeeping model incorporates elements of the inner-sense theory of self-knowledge.

Contra Schwengerer, I don't think there is any use in trying to fit the double bookkeeping account either into the inner-sense or the introspective transparency paradigm. The account agrees with transparency theorists in the sense that what introspective inferences take the content of other world-directed states as premisses. This explains the afore mentioned fact that there is no special introspective phenomenology because the appearance properties involved in introspection will be the ones involved in ordinary perceptual inference. But also, the relevant inferences are inductive and fallible and can thus me likened to perceptual inferences in this sense.

Hohwy's model suggests a straightforward extension to account for introspective attention. If attention is explained by expected precision then introspective attention

[12] William Lycan for instance is committed to a limited transparency claim while also defending an inner-sense theory of self-knowledge (Lycan 1996).

[13] As Dretske says, "[i]f [transparent self-knowledge] is inferential knowledge, it is a strange case of inference: the premises do not have to be true to establish the conclusion." (Dretske 1995, p.61). In my estimation our account of self-knowledge is immune to this problem of giving a coherent justification for transparent inference, which has been called the "puzzle of transparency" (Barz 2019). Of course, were we to conceive of inference in terms of logical deduction and included the transparency move $P \vdash \mathit{Represent}(P)$ this would yield havoc. For this would also enable one to conclude $\neg\mathit{Represent}(P) \vdash \neg P$. But if we think of the relevant inferences as Bayesian inductions no such issue arises.

is explained by increased expected precision for predictions of mental causes. Then even the smallest mismatch of prediction and fact will start to drive hypothesis revision. We will later discuss the interaction between introspective attention and predictive dynamics in more detail. Note that this view on introspective attention fits nicely with our model of the hierarchical analysis of intentional objects into appearance properties in the previous chapter.

The double bookkeeping model further suggests that such introspection mostly occurs when prediction error cannot be explained away in terms of worldly causes alone. Thus in a normal course of action introspection will be unnecessary. This aligns nicely with the post-Husserlian phenomenological tradition that tends to emphasise that humans normally find themselves in a direct relationship with their world where introspection and the separation of the world into a representing subject and an independently existing object are the exception rather than the norm.[14] Both phenomenological analysis and theory-based reflection suggest that introspection is typically a result of the breakdown of the more original unmediated mind-world relation or, of course, of a deliberate attempt to know one's own mental states. But also note that perceptual and introspective inference will typically not be neatly separated in the sense that either one or the other is singularly operative at every point in time. Rather, in different scenarios the processing of information will be more or less reflexive, i.e. introspective hypothesis will be more or less relevant in explaining away prediction error.

An interesting feature of the double bookkeeping account of consciousness is that it holds the potential of explaining the apparent difference in access to our dispositional and occurrent mental states. While we have not provided an exact mapping of folk-psychological notions of dispositional propositional attitudes it is cogent to hold that such attitudes will be encoded in the generative model of a system. On the other hand, occurrent mental states will be encoded in current predictions. This immediately entails that dispositional and occurrent mental states cannot be meaningfully fully disentangled. We saw in chapter six that the causal structure attributed to the world by some prediction will be determined by the generative model. On the other

[14]This view is of course most famously defended in Heidegger 1927.

hand, dispositional mental states only enter into our mental life insofar as they shape predictive dynamics. But still we can draw a meaningful principled distinction here.

If the double bookkeeping model is correct then we have a straightforward explanation of why dispositional mental states are harder to introspect than occurrent ones, even though both are ultimately encoded in the dynamics of the predictive hierarchy. That is because predictive processing systems do not explore significant parts of their overall generative model at any given time. Predictions are used in approximate Bayesian inference precisely to limit the extend to which one has to explore epistemic possibility space. Thus it will generally be a hard inference task to infer one's dispositions towards environmental conditions that one has not seen before from those one has been in before. On the other hand, inferring the state of one's lower-level representations, i.e. one's current prediction, is a trivial task in comparison.

Representations pertaining to one's own dispositional mental states are invaluable in planning one's action because in doing so oneself is just another cause determining outcomes. It has been proposed that the nervous system's attempt to model it's own generative model as a subset of the world is the basis of the self as an enacted model.[15] But following these speculations would lead us too far afield.

Let us now return to our subject matter of phenomenal self-knowledge. It is plausible that the double bookkeeping model may explain phenomenal self-knowledge as a special case. This is particularly plausible on our representationalist point of view that holds that conscious content is a form of predictive content. As the content of predictions can be directly inferred based on prior knowledge and prediction errors from the relevant layer of the hierarchy we have a powerful model at hand for explaining self-knowledge. Let us now investigate how this proposal illuminates the reflexive nature of consciousness.

8.2 The Dual Nature of Consciousness

The reflexivity of consciousness denotes its characteristic of presenting a content to a subject. It seems impossible for there to be a phenomenal experience of blue, say, without someone experiencing the quality as presented to oneself. We have

[15] Hohwy 2014 chapter 12.

tentatively explained this phenomenological feature of experience in terms of its reflexive representational properties. The relevant conscious experience of blue, we hypothesized, not merely represents a primitive appearance, but it also represents that it so represents. So to integrate the reflexive nature of consciousness into our model we would have to explain self-representation in purely inferential-computational terms.

Explaining the self-representational nature of consciousness is not merely necessary to explain its apparent reflexivity, but, as we saw in chapter two, it is also invaluable to understanding how it is that conscious experience represents appearance properties. Appearance properties are primitive ways of appearing. Above we tentatively explained how the predictive hierarchy comes to represent the world in terms of semantic primitives encoded at low-levels of the hierarchy. But to explain the role of appearance properties in consciousness fully we would have to understand how the inferential hierarchy represents appearance properties semantic primitives, i.e. *as* ways of being represented. In other words, so far we lack an explanation of the self-representational nature of low-level representations in the predictive hierarchy.

Now it is natural to hold that the double bookkeeping model of self-knowledge straightforwardly explains the reflexive nature of some low-level representations. In our inferentialist paradigm we may say that the duality of qualitative and reflexive consciousness is explained by an underlying duality of *object-oriented* and *subject-oriented* inferential relations. Object-oriented inferential relations are those in virtue of which an environmental cause is represented. Subject-oriented inferential relations are those in virtue of which the object is represented *as being represented*.

I see a blue rose. Thus the appearance property of blue, represented by some specific low-level representational state, is represented as being caused by an environmental cause, a blue rose. All this is explained by the object-oriented inferential role of low-level representations expounded on in detail in the previous chapter.

But the property of appearance blue, our analysis in chapter two suggested, just is a way of being represented. To explain how some state represents an appearance property it is not enough to explain how it represents a primitive property. One also has to explain how it is that this property is represented as a way of being represented. On the current model this is explained by the subject-oriented inferential role of the low-level representational state represented that is explained in terms of the

double bookkeeping model. The subject-oriented inferential role is constituted by an inferential relation to a higher-level representation that pertains to the semantic properties of the low-level representational state.

At this point some readers may worry that I am misframeing the described predictive dynamics when I hold them to be an instance of self-representation. Isn't it the case that reflexive consciousness is not explained in terms of one representation representing itself, but in terms of a higher-order meta-representation of the first-order state? This would entail that the theory at issue is really a *higher-order theory of consciousness*, a theory of consciousness that holds that consciousness is essentially meta-represented by distinct higher-order representations.[16]

In my view there is no genuine problem here. Remember that inferentialism is holistic such that representations cannot be neatly separated into singular states without loosing all their representational import. This applies to representations and higher-order representations. There is no neat fact of the matter whether a first-order state and a second-order state are to be considered two states or as aspects of one inferential network. After all, the second-order state would not be a second-order state if it would not bear its particular inferential relation to the first-order state. Thus on the inferentialist account of representational content there is no strong distinction between higher-order theories and self-representationalism.[17] I will discuss some of the apparent problems that may result from a disentangling of object-oriented and subject-oriented inferential role below.

Now for the first time we can formulate what consciousness really is according to Bayesian representationalism: Consciousness is an inferentially grounded reflexive representation that involves semantic primitives. It arises when complex computational systems perform hierarchical approximate Bayesian inference about the causes of sensory stimulation. Such representations will naturally turn out reflexive in order to correct for prediction errors arising from causes within the inferential hierarchy itself. However this suggestion leads to a rather obvious problem: Re-

[16] Rosenthal 1986.

[17] The resulting view has some similarities to the so-called *wide intrinsicality view* according to which conscious states and higher-order states are parts of one and the same mental state (Gennaro 2006). In the context of inferentialism the mereological claim implicit in the wide intrinsicality view is true in virtue of inferential structure. However, I disagree with Gennaro that the view necessarily excludes scepticism about one's own experiential states (Gennaro 2006, p. 225).

flexive and qualitative consciousness should be able to exist on their own because they are realized by distinct inferential-computational processes. This seems *prima facie* implausible. Let us discuss whether it really is.

8.3 An Ambiguity of 'Consciousness'

Ordinary consciousness is both reflexive and qualitative. These aspects of consciousness we argued however, are independently realized by different inferential roles. These independent inferential roles are just different extrinsic functional features of one and the same state within the predictive hierarchy. Thus it should at least be possible if not sometimes actual that they come apart. There should thus be states that bear a subject-oriented inferential role without bearing an object-oriented one and *vice versa*. This seems to entail the unsettling and phenomenologically puzzling consequence, namely the disentanglement of qualitative and reflexive consciousness. What ought we to say about such prospects?

Orthodox self-representationalists will be prone to hold that only properly self-representational states are conscious. On such a view only a state that bears a dual inferential role is a conscious state strictly speaking. If one falls away the state becomes unconscious. The story I want to put fourth is more complex. I will argue that the possible cases of purely qualitative or purely reflexive consciousness actually underlie ambiguous psychological states that in some sense are conscious and in another sense are not. Thus I hold that our ordinary notion of consciousness is not be fine-grained enough to capture the full scope of psychological complexity.

First let us discuss cases where one state bear a subject-oriented inferential role without bearing an object-oriented one. Arguably, this is best described as a form of introspective illusion. As a result of faulty introspective inference some state is represented to represent some state of affairs but really it does not so represent. How common are introspective illusions of this kind?

It seems that we can interpret Schwitzgebel's insight that introspection of conscious experiences is a much more elusive business than one may intuit as evidence for the kind of introspective illusion we are looking for. There are many presuppositions we hold about experience that seem to crumble as soon as we interrogate them in detail.

For instance, while at first it seems that we intrinsically know the phenomenology of emotion, you will realize that pinning it down in detail is a much harder task than one may think at first.[18] I for one had the intuition that emotional phenomenology is not clearly localized in the body. But when taking up meditative practice it became obvious that this was arguably a kind of introspective illusion: Emotional phenomenology can almost universally be localized in specific parts of the body.

What seem to be introspective errors can be nicely explained as breakdowns of introspective inference. In other words we are dealing with states that bear a certain subject-oriented inferential role without bearing the correct object-oriented inferential role. So far the possible disentanglement of qualitative and reflexive consciousness seems to be phenomenologically unproblematic.

An interesting feature of thinking about introspection in this inductive fashion is that it will entail a form of context-dependence of introspection that we would normally deny exists in knowing our mental states. Reconsider the example of inferring that one is dizzy from the world's spinning. Here the background knowledge that one normally does not just start rotating implicitly facilitates introspective inference. I can only speculate that the mistake of thinking of emotional phenomenology as non-localized may be grounded in an overly Cartesian view of the mind where mind and body are quite distinct. So while the double bookkeeping model entails a privileged access to one's mental states it does so in a way that is potentially sensitive to socio-culturally engrained background assumptions impacting our capacity of self-knowledge. Our inner lives are not just transparently given in a naively Cartesian sense. Just as the perceptual realm the introspective realm may be subject to prejudice.

Returning to the issue of the phenomenological plausibility of a Bayesian rendering of self-representation, I suspect that many readers will resists my account of the independence of qualitative and reflexive features of consciousness. While one may grant that one sometimes makes introspective mistakes based on inattention and wrong phenomenological presumptions, at least at the center of vision and where appearance properties are concerned such mistakes are out of question. And indeed it

[18] Schwitzgebel 2006.

is hard to imagine that, when I attend to the appearance of the sky directly, I may be subject to introspective illusion and the sky really appears some wholly different way.

If introspective inference is inductive it must be fallible.[19] So if our introspective intuition is one of infallibility then our intuition goes astray if the current model of self-knowledge holds any water. However, the current account *can* explain how our intuition of security for these states arises. Low-level representations usually will be in a tightly constrained set of possible states. For instance, certain low-level representations will represent appearance-colors. It is natural to hold that inferences where only a limited amount of possible outcomes are at issue are very reliable. The situation would maybe be very different of we were frequently presented with genuine new colors that have no place in the ordinary color-space. In such a situation the proposition that phenomenal mental states, even at the center of attention, are easily knowable may seem dubious. But as it stands we can straightforwardly explain how it comes that low-level representations appear easily knowable (i.e. with high precision).

I conclude that our Bayesian account of reflexive consciousness cannot obviously be rejected on phenomenological grounds. Let us now turn to the more interesting case of states that bear an object-oriented inferential role without bearing a subject-oriented one. This should lead to cases where qualitative consciousness occurs without associated reflexive consciousness.

First of all, note that, unlike cases of reflexive consciousness without qualitative consciousness, the reverse case does not entail that one is in any sense deluded. Rather something that is represented is not represented to be represented. We may express this by saying that the representational system as not made it's own representational states explicit *as* representations. But how may such cases be reflected in phenomenology?

I want to suggest that cases of absent-mindedness are what we are looking for. Imagine a truck driver driving his route from Frankfurt to Hamburg for hundredth time. On arrival he suddenly realizes that he has been driving 'without thinking about it' for the whole time and is surprised that he has reached his destination already.[20]

[19] More generally, every account that holds that a mental state and the knowledge of that mental state are independently realized entities should also hold that self-knowledge is fallible (Armstrong 1968; Barz 2021).

[20] The example of the long distance truck driver goes back to Armstrong 1980.

On the account I want to suggest we should describe this case in the following manner. In some sense it would obviously be false to hold that the truck driver has been driving unconsciously! The man wasn't asleep or comatose. He saw the other traffic participants and street signs and reacted to them adaptively. So there is some sense in which he was conscious the entire time - conscious of the cars and the street signs. In other words, he enjoyed qualitative consciousness, that is consciousness of features of his environment.

But at the same time the truck driver wasn't conscious in another sense and only became conscious when he arrived at his destination and wondered how he could have driven the whole route as though in trance. It should seem natural to hold that this other sense is that of reflexive consciousness. The truck driver was qualitatively conscious of features of his environment but he wasn't reflexively conscious of these very qualitatively conscious states. While he enjoyed full-fledged representations of the external world, these representations have not been made explicit as representations. We will return to the issue of differentiating between conscious and unconscious representations in quite some detail below.

We can straightforwardly explain why the truck driver wasn't reflexively conscious by remembering how reflexivity arises in the predictive hierarchy. Reflexive hypotheses become important in explaining away prediction error only where it cannot be explained away in a world-directed manner. So in highly automatized tasks where no strange unexpected stimuli arise we should expect reflexive representations to be less important in the inferential process. As Erwin Schrödinger was early to notice, consciousness seems to be associated primarily with the *newness* of a certain task that cannot be handled by our unconscious automatisms.[21] We can capture this observation that consciousness seems to be bound up with newness by holding that reflexive consciousness typically arises where introspective predictions become necessary in explaining away unusual prediction errors.

If the current analysis is sound then there arguably is no answer to the question whether the truck driver *really* was conscious prior to arriving at his destination. Rather the idea here is that our common sense concept of consciousness ambiguous

[21] Schrödinger 1958.

between two senses or three senses. One is qualitative consciousness, one is reflexive consciousness, and one is a *fully conscious* state that includes both aspects.

In chapter two we committed to self-representationalism on the grounds that the content of representations of appearance properties cannot be expressed without recourse to the representational state itself. We must now qualify this commitment. Self-representationalism is true in the sense that fully conscious states are states that both represent an object in some way and they represent that they so represent. Self-representationalism however is false in that these aspects can be meaningfully disentangled and the remaining states in some sense fall under our folk-psychological concept 'consciousness'. For the truck driver was conscious the entire drive.

Here a tangential question arises. Should we hold that beings who's representations only bear object-oriented inferential roles and can't even potentially reflexively represent their own mental states be counted as conscious beings? Is at least the potential for reflection necessary here? My hunch is that it is misguided to search for a yes-or-no answer. Just as the truck driver such agents will be conscious in one sense while not being capable of consciousness in another. Nature is under no obligation to be neatly sort into our predefined categories.

We made some leeway in our understanding of consciousness in terms of the inferential structure of brain and mind. But we are still far from a complete understanding. We still lack a principled account of why certain mental processes happen in the light of consciousness while other are in the dark. So far the account has nothing to say about the psychological differences between such states. I will now try alleviate this flaw.

8.4 The Conscious and Unconscious

The problem we want to tackle in this section is that so far we haven't given an account of the divide between conscious and unconscious representations. Unconscious representational states are theoretical postulates that can take a variety of forms. Psychodynamic theories postulate unconscious beliefs, desires and thoughts that are unconscious due to a process of repression. Perceptual psychology shows that there are perception-like processes that do not enter conscious awareness, i.e. cases of subliminal perception. And of course cognitive science postulates unconscious

representational states like Chomsky's deep grammar that are supposed to explain mental functioning.[22] In this latter vein, the predictions postulated by predictive processing, the particular cognitive theory we are interested in, are not conscious all the time. Arguably the hierarchy tracks much more possible states of the world that we are conscious of at any given time. It seems as if we are primarily conscious of a final synthesis of the predictive dynamics rather than of the whole state of the world, as it is currently predicted. This entails a challenge to Bayesian representationalism: If predictions aren't by their very nature conscious, then we need to explain why some representations are conscious, ideally without appealing to anything beyond inferential processes. In this section we will engage these tasks.

We will proceed in three steps. First we will briefly discuss one of the more promising approaches to consciousness in the neurosciences and the evidence for it, namely the global neuronal workspace theory. We will then explain how the global neuronal workspace theory can be explained as an emergent feature of the predictive dynamics of an agent performing active inference. This will entail a model of the psychological dynamics of consciousness and a model of the conscious-unconscious divide.

8.4.1 The Global Neuronal Workspace

The *global neuronal workspace theory* of consciousness is one of the most well developed and experimentally sound scientific accounts of the neuronal basis of consciousness we possess to date.[23] It arose in the context of *modular theories of mind* that hold that the brain can be subdivided into a variety of different functionally isolated modules that are each specialized on different kinds of information processing.[24] These modules are imagined to, by themselves, process information unconsciously. This fits nicely with the psychological finding that almost all kinds of information processing that we engage in consciously can be also be accomplished unconsciously, i.e. they aren't associated with introspectable phenomenal character. From semantic (i.e. meaning-sensitive) processing of language to complex visual processing, it seems that no conscious states are strictly required in any of these.[25]

[22] Cowie 2008.
[23] The idea was first introduced in Baars 1988. For a recent review, see Mashour et al. 2020. For a popular introduction, see Dehaene 2014.
[24] Dehaene and Naccache 2001.
[25] A nice early summary of these findings can be found in Velmans 1991.

But if all tasks we usually engage in can be accomplished by unconscious cognitive modules, why do we possess consciousness at all? Here the global workspace comes in. The cognitive modules cannot work productively without sharing information. The idea is that, in some sense, they must be 'on the same page' with regards to their representations of the external world and what is desirable and important at any given time. And this is precisely what the global workspace achieves. It is thought to be a global neuronal network that is supposed to establish an information-sharing space accessible by all cognitive modules such as to make coherent action possible. The information that is made globally available in this way is thought to constitute the content of consciousness.[26]

Neuroscientific support for this framework is found in the phenomenon of *ignition*. This is a globally increased neuronal activity observable when a certain content crosses into consciousness. To observe this experimentally one may show a subject stimuli of varied intensity. At some point they will be so intense as to be consciously perceptible. As it turns out, precisely at this point one can observe an 'avalanche' of neuronal activity across different parts of the cortex.[27]

Interestingly, processing that is not associated with consciousness is not necessarily short-lived or low in intensity. However it is generally associated with decaying waves of activity, rather than the self-reinforcing phenomenon of ignition involving disparate brain areas. As one may expect, given the hypothesis that ignition is associated with a content entering the global workspace, the factors that facilitate ignition are thought to be firstly stimulus intensity and secondly attention to the stimulus.

In line with the global neuronal workspace theory Stanislas Dehaene has hypothesised that the two observable modes of functioning, one unconscious and not associated with decaying local activity, one conscious and associated with ignition, should be interpreted as different modes of evidence accumulation by the brain. In one mode it slowly gathers evidence, changing is representations about the world incrementally. In the other, in the 'conscious mode' the brain integrates evidence gathered by different cognitive modules, thereby making large changes in internal

[26]Dehaene and Naccache 2001.
[27]Mashour et al. 2020.

representations possible.[28] From the standpoint of the predictive mind this must seem suggestive as these sound like different kinds of predictive dynamics.

Original theories of the global workspace were formulated in a thoroughly bottom-up processing paradigm. But as the theory is both theoretically well motivated and empirically supported it would be interesting to see how it may be reframed within the predictive paradigm. We will now show that, by considering the central role of active inference, we can explain the emergence of a global workspace as an emergent feature of the predictive mind.

8.4.2 The Winning Hypothesis Account

In this subsection I will show how we can explain the existence of something like a global neuronal workspace as an emergent[29] feature of a predictive brain. The central idea here is that active inference necessitates that the brain settles for a *winning hypothesis* in the guidance of action. The winning hypothesis will consist of the global hypothesis about the state of the environment that most minimizes expected prediction error. Because this requires an integration of information of all parts of the predictive hierarchy, this process is functionally equivalent to the sharing of information within a global neuronal workspace.

The best way to approach these ideas is first considering an agent that is not capable of action and therefore will not settle for a winning hypothesis. Imagine for instance Galen Strawson's weather-watcher thought experiment. These are alien beings that evolved on some world orbiting a distant star. To any human visitor these critters will seem like nothing more than strange trees. But through some fluke of evolution the weather-watchers have an interesting skill. Their internal network of fluid transporting tubes resemble an animal nervous system. When rain falls into the funnels at their top they register the internal network processes this information in a perception-like fashion. When wind bends their thick trunks, this is registered, too. The funnels and the trunks serve as the weather-watchers sense organs. And for some reason their fluid-based nervous system engages in predicting the weather to an accuracy that man could only achieve using satellite imagery and computation devices.

[28] Dehaene 2011.

[29] I use the term in the sense of weak emergence, i.e. the workspace is a seemingly novel property that in principle can be predicted from the properties of the constituents.

Aside the biological oddities, the philosophically interesting thing about whether-watcher is that they perceive though they do not act. They don't even conceive of possible actions. All they are are highly complex and, as far as it goes, intelligent *pure perceivers*.[30] Let us now further stipulate that the weather-watchers employ a predictive hierarchy to perceive. Evidently, for optimal predictive performance, weather-watchers should not predict the local atmosphere to be in any particular state determinately. Rather, based on the limited information flowing in from their trunks and funnels, they should represent various states of the world with various probabilities. Thus, here it is hard to see how the weather-watchers could pick out any singular hypotheses as *the* state they perceive their environment to be in right now. Their predictive task will be fulfilled more fully by representing a non-peaked probability distribution over various states.

How are we different from weather-watchers? We act. Imagine grasping for your coffee cup. Remember that in the active inference paradigm there is no strong distinction between desires and beliefs. Action is initialized by being expected, we may say. Now suppose that, based on the limited information you possess, that the cup is represented as a probabilistic blur in various locations. Here, when trying to grasp the cup, an obvious dilemma arises: You can't grasp a probabilistic blur. That is, if you were to represent your environment similarly to a weather-watcher you should arguably grasp the cup in a variety of locations with a variety of probabilities. But in action an important symmetry break occurs in that you can only grasp the cup in one of its possible locations.

We can now understand how it comes that our predictive hierarchy seems to synthesise its activity into a single winning hypothesis. The system needs to select *one* hypothesis about the state of the world act upon it, based on their estimated posterior probability.[31] Active inference seems to require that there occurs a *probabilistic collapse* of the represented probability distribution over current states of the world into a sharply peaked distribution.

We can think of probabilistic collapse as initialized by a strong prior belief that at t_0, i.e. now, the probability distribution will be sharply peaked when acting. This entails that if one holds the world to be in a variety of states with

[30] Strawson 1996. Note that my description differs from Strawson's in astrobiological details.
[31] Hohwy 2014; Hohwy 2012.

different probabilities this will increase prediction error thereby facilitating the elimination of low-posterior hypotheses.

So far the winning hypothesis model is a relatively straightforward derivation of the active inference view of the relation of action and perception. Still, it is a powerful tool to connect Bayesian cognitive science both to the phenomenology of consciousness and its neurophysiological correlates. To achieve this it is necessary to see that *reporting* on your inner state is a kind of active inference. Thus such reports may initiate probabilistic collapse or in other words, we can only ever introspect and report on a winning hypothesis. Trying to see inside to tell what one is conscious of determines one's mental state as much as is makes preexisting states available for report.

This first of all answers Block's worries about Bayesian views of perception. While we saw in the previous chapter that consciousness is in some sense more probabilistic than it may seem at a first glance, and secondly, that out of the box predictive processing can explain that perception does not represent a probability distribution about the totality of epistemic space, there remained an impression that the Bayesian view is not wholly true to perceptual phenomenology. When we introspect our perceptual states it appears that they represent the world to be in one specific state rather than a variety of states. We now know why. Introspection, as a kind of action or active inference, collapses the probabilities tracked by the brain into a single winning hypothesis, thereby suppressing the introspection of countervailing hypotheses. That perceptual representations almost never seem probabilistic is partly an artifact of how we come to know them.

A further phenomenological feature explained by the winning hypothesis account is the fact that introspective elusiveness as described by Schwitzgebel is less pervasive for simple features like appearance properties but more so for higher-level features. Introspecting the precise phenomenal character of a red experience is easy, introspecting the precise phenomenal character of a feeling caused by a work of art is much less straightforward.

If what we are trying to introspect is really a winning hypothesis such a phenomenon is to be expected. Currently available sensory states serve as the boundary conditions of introspective inference. All possible hypotheses that have a chance of 'winning' will have to explain away this incoming information. Thus the rele-

vant states will be relatively stable and the introspective inference will be a low-ambiguity inference associated with high precision.[32]

On the other hand, introspecting the higher-level features of a winning hypothesis will be much more tricky because this hypothesis will be ever shifting on the way to figuring out the best possible hypothesis about the environment. Introspective inference of higher-level representations will thus naturally be associated with lower levels of expected precision. One's inner life is experienced as elusive and hard to pin down[33] while our direct experience seems just directly available.

Beyond phenomenology the winning hypothesis account helps us to connect the idea of a globally occurring neuronal workspace with our Bayesian account of cognition. The settling for a winning hypothesis will be a global process involving the synthesis of contents from all over the hierarchy into a globally optimal prediction. This entails that probabilistic collapse should be associated with global neuronal activity. Thus we have a good candidate for a principled account of what the global neuronal workspace is. It is an emergent feature of an hierarchically realized active inference system.

This proposal of course makes good sense of the phenomenon of ignition. First of all, ignition is associated with attention. In our predictive paradigm attention is associated with precision and high precision entails low expected prediction error. As the winning hypothesis will be the one minimizing expected prediction error this entails that stimuli that are attended to will enter consciousness and entail ignition with greater propensity.

Ignition was further associated with stimulus strength. Of course, stimulus strength is not a variable that any Bayesian system will be directly sensitive to. However, it is reasonable that stimulus strength roughly equals the stimulus being *stronger than expected*. Thus stimulus strength will be strongly associated with higher prediction error and thus has to be explained away by the winning hypothesis.

These ideas cohere so naturally that they have been combined into a unified model called the *predictive global neuronal workspace*. This model seems to have some empirical advantages over the original global workspace approach as it accounts for the fact that ignition seems to be sensitive to expected uncertainty in some context, which fits nicely with predictive approach. In fact this is precisely what we expected: Diver-

[32] Clark, Friston, and Wilkinson 2019.
[33] This hypothesis is inspired by Hohwy 2014, p. 247-248.

gence from expectation rather than mere stimulus strength facilitate ignition.[34] While these developments are new this seems to be a very promising area of further research.

The most important contribution that the winning hypothesis model has to contribute to Bayesian representationalism is a potential principled understanding of why some representational states are conscious while others are not. Here the suggestion would be that we experience mental representations as conscious states if they are integrated into the winning hypothesis. However we will see in the following section that there are good reasons not to commit to this thesis fully.

8.5 The Amalgamation of Consciousness

The idea of a predictive global neuronal workspace holds the potential of a unified theory of consciousness in the context of Bayesian cognitive science. However not all is well. First, the attentive reader will have noticed that we by now have produced two distinct accounts of how some content becomes conscious, one based on the coming together of qualitative and reflexive features, one based on its integration into the winning hypothesis or just *integration* for short. How is it these accounts can be true at the same time?

Secondly, the winning hypothesis account seems to struggle to explain certain kinds of unconscious processes and their relation to consciousness. In so-called *blindsight* subjects that seems to be blind in a certain part of their visual field seem to be able to guess the content of this part of the visual field extremely well.[35] It seems that information from this part of the visual field can influence action without entering consciousness. This entails that being recruited for action guidance and being conscious cannot be one and the same thing. This seems to put the winning hypothesis model and its underlying representationalism in jeopardy.[36]

In case of the latter problem I suggest that we should explain unconscious influences on action in blindsight subjects as resulting from low-precision input from the visual system. As blindsight is usually caused by lesions[37] we may speculate that such low-precision is the result of the breaking away of large chunks of information flow

[34]Whyte 2019; Whyte and Smith 2020.
[35]Weiskrantz 2007.
[36]Dołega and Dewhurst 2021; Marvan and Havlík 2021.
[37]Weiskrantz 2007.

from impacted areas. As, compared to the previous state, the relevant part of the visual field is virtually blind, precision will be estimated to be very low. Such low-precision input will not be sufficient to ignite the global workspace[38] because the relevant prediction error is too weak to require active inference for a resolution. Still, if the agents acts the relevant prediction error will have some influence on the shape of the approximate posterior. Thus the relevant content will have entered the global workspace even if, by itself, it would not be capable of igniting it. Why then isn't the visual input in blindsight subjects conscious?

Here we can appeal to the ideas expounded above. Consciousness is characterized by a coming together of introspective and perceptual inference. Low-precision input, like the input from the blind-sighted visual field will generally entail low prediction error. Low prediction error entails that, even though some state may have been integrated into the winning hypothesis, the relevant state will not be represented *as* represented and will thus lack full phenomenal consciousness. In a nutshell, low-precision input can be expected to be introspectively opaque.

If being phenomenally conscious is ultimately explained in terms of self-representation, what role is there to play for the winning hypothesis account? The answer is that self-representationalism explains the *metaphysical nature of conscious content*. The winning hypothesis model on the other hand explains the *psychological role of conscious content* in the predictive mind. This means that representations that are integrated into the winning hypothesis are not by their very nature conscious. Only self-representational states are. However being self-representational is strongly correlated with integration. Thus which states do and do not become conscious is often best explained with recourse to the current winning hypothesis.

The correlation between self-representation and integration has at least two aspects. First, stimuli that are likely to be ignite the global workspace and thus that the winning hypothesis is trying to explain away is likely to also entail introspective inference. This is because we must be dealing with a stimulus that is high in prediction error, high enough indeed to necessitate active inference, which in turn necessitates probabilistic collapse. As we saw above, high prediction error will increase the propensity for introspective inference because it may be the result of an error internal

[38]Derrien et al. 2022.

to the hierarchy. For instance, a loud and unexpected sound will both be precisely the kind of thing that provokes conscious awareness and that provokes introspective inference to ensure whether the stimulus has a genuinely exogenous source.

The second aspect of the connection between integration and self-representation is the already discussed fact that introspective inference may entail probabilistic collapse. Thus deliberate introspection, which is one way of making a state reflexive, also entails integration. If I start attending to my current mental state, maybe to report on it, this will itself entail settling for a winning hypothesis to report on.

So taken together Bayesian (self-)representationalism and the winning hypothesis model yield a promising account of the nature and the cognitive role of consciousness. Consciousness is a certain form of reflexive representation of the causes of sensory stimulation. Such representations emerge either due to exogenous stimuli that both entail introspective inference and integrated active inference or due to endogenous active inferences to report on mental content. Both are associated with probabilistic collapse and ignition.

On first approximation it is natural to hold that consciousness is a metaphysically simple. It is just an unanalyzable inner glow. This naturally leads to the suspicion that consciousness will have a distinct and unified metaphysical nature, i.e. it will turn out to be natural kind. The current investigation indicates that this is not the case and that what we ordinarily refer to as consciousness rather is an amalgamation of metaphysically distinct features of mental representations. Above we already saw that qualitative and reflexive consciousness can be meaningfully disentangled. Now we see that there is a further feature, integration, that explains further introspectable features of experience that can also be distinguished.

The most prominent rival of the global workspace theory is integrated information theory, which we argued has trouble accounting for the representational nature of consciousness. Remember that integrated information theory posits that consciousness is connected to the recursion within an information processing system. The intuition here is that consciousness is in some sense intrinsically self-directed. On the current account this intuition is vindicated by the admission that fully conscious states are by their very nature reflexive. However, other than integrated information theory, I explained this intuition in terms of representational rather than vehicle properties.

In the following section I will show that there are a few interesting predictions about the nature of the unconscious to be drawn from all this.

8.6 Note on the Unconscious

While it is not at the center of our concern the ideas just elaborated have interesting consequences for the ontology of the unconscious. Unconscious mental states can generally be separated into dispositional and occurrent states. I already commented on the fact that dispositional states are quite hard to infer and thus will be unconscious generally speaking. Here we will focus on occurrent unconscious states, i.e. representational states that are not associated with introspectable phenomenal properties.

If the previous self-representationalist synthesis is correct then unconscious mental states will be hypotheses that fail to be integrated into the winning hypothesis that constitutes the content of the global workspace, or states that are so integrated without being introspectable. This will generally mean that the hypothesis is expected to reduce prediction error less than some alternative hypothesis. This may be the result of low precision, low prior probability or low likelihood or a combination of these.

While there has been some interest in the revival of psychoanalytic idea in the light of the Bayesian paradigm,[39] the winning hypothesis model has the potential to illuminate this issues to a higher degree than has been noticed so far. In particular it may provide a neuro-cognitive basis for some of the most central concepts of psychoanalytic theory. To see this, let us just review three ideas central to psychoanalytic thinking, namely first the mechanism of repression, secondly the idea of over-determination and thirdly, the idea of the unconscious as compensatory.

According to psychoanalytic thinking the normal mechanism for rendering a certain mental state unconscious is *repression*. Repression is best though of as a process of sub-personal self-denial of mental contents that conflict with one's self-image and/or social norms. From our Bayesian perspective we can render this in terms of a state's prior probability. On the active inference account states that conflict with social norms or one's self-image will be associated with low prior probability. They would thus considerably raise the expected prediction error if they were

[39] For an investigation of the qualitative agreements of the structure of psychoanalytic thinking and Bayesian cognitive science, see Carhart-Harris and Friston 2010 and Solms 2021.

taken to be true! Thus repression of conflicting ideas is a natural consequence of the winning hypothesis account. These mental contents will subsequently be denied entrance to the light of consciousness and may only be revealed by intense introspection, maybe facilitated through therapy. A natural and probably testable prediction of this view is that mental processes subject to repression will be associated with decaying waves of neuronal activity.

Another psychoanalytic idea that is resonant with the view on the unconscious defended here is that of *over-determination*. According to Freud the images in dreams or Freudian slips are not necessarily to be interpreted in a single way but can be symbols for a number of different unconscious contents at a time.[40] While it is legitimate to speculate that here we are dealing with the attempt of immunizing a theory against falsification, there is also a plausible interpretation of these ideas in probabilistic terms. As we conceive of the unconscious as the 'storehouse of discarded hypothesis' it will, by its very nature, represent many contradictory states of affairs at a time without settling for a single unified hypothesis. Internal coherence and unity is a distinctive feature of consciousness as opposed to the unconscious.

Finally, in analytical psychology the Freudian idea of repression has been developed into the concept of *compensation*. Here the idea is that the unconscious will, as a general rule, contain contents that are diametrically opposed to conscious ones. This can be nicely illustrated by cases where one may consciously adopt a certain position and then have dreams that deal with ideas that oppose these positions.[41] In analytical psychology this observed tendency is explained by appeal to an underlying wholeness of the that is then divided into conscious and unconscious aspects. But on the Bayesian picture there is of course a much more straightforward explanation: The selection of a certain winning hypothesis entails that rival hypotheses are suppressed. This however does not mean that these hypotheses are completely discarded. Far from it they may, while being plausible, be discarded based on their radical incompatibility with the winning hypothesis while retaining a moderate posterior probability otherwise.

These are preliminary but, I presume, natural suggestions. Developing them here would seriously strain the constraints of our task which is, after all, the nature of our conscious mental life rather than its dark mirror image. I hope that

[40] Freud 1966, p.239.
[41] Jung 1971, p. 110 f.

by now I have convinced the reader of the fact that the predictive processing account may serve to explain many of the features of conscious experience and its relation to the brain and maybe the unconscious. However so far traditional issues of the metaphysics of consciousness have been curiously absent from my discussion. Let us now engage with these issues.

8.7 The Hard Problem

The hard problem of consciousness is the problem of explaining how it is that physico-functional states are associated with phenomenal consciousness at all.[42] The problem is best illustrated by *conceivability arguments*. For instance, I can imagine a *philosophical zombie*, a being that is physio-functionally just like me but not conscious. Thus it seems hard to see how the physico-functional facts about me could explain the phenomenal facts about me.

Has Bayesian representationalism anything to offer when it comes to solving the hard problem as evinced by the conceivability argument? Arguably it does not. For it seems that we will be able to imagine a zombie quite independently of which specific physicalist theory we assume to capture the nature of consciousness. Bayesian representationalism seemingly cannot explain the difference between zombies and non-zombies.

There is another family of arguments that can be used to show that consciousness poses a hard problem. Defenders of the *knowledge arguments* deduce from the premise that it seems that no amount of physical knowledge allows one to derive phenomenal facts *a priori*, phenomenal facts are non-physical. The canonical version of this argument was delivered in the form of Frank Jackson's Mary argument: Mary is a genius neuroscientist who knows everything there is to know about the physics color vision. However, she has lived her whole life in a black and white room. When one day she leaves the room, she sees colors for the first time. She then learns something about color vision, namely what it feels like. Thus all the physical facts were insufficient to know the relevant phenomenal facts. Thus phenomenal facts are non-physical facts.[43]

[42] Chalmers 1995.
[43] Jackson 1986.

Has Bayesian representationalism anything to offer when it comes to solving the hard problem as evinced by the knowledge argument? On the first glance it hasn't. For even if Mary did know that appearance colors are represented by low-level representations in the predictive hierarchy, this would not allow her to conclude what it feels like to see blue, say.

However, the situation is a little more complex. For while it is true that Mary cannot derive knowledge of the relevant experiential quality *a priori*, Bayesian representationalism will not leave her completely in the dark either. For, given that she knows that she will have a new kind of experience representing a novel *sui generis* semantic primitive, she will be able to deduce *that* she will learn something about an experiential quality, even though she is still in the dark regarding what this quality will feel like. We can capture this by saying that Mary lacks *first-order knowledge* (knowledge about the specific experiential quality), she possesses *second-order knowledge* (knowledge that there will be such a quality).

All in all it seems that Bayesian representationalism doesn't offer a huge help in solving the hard problem. To provide such a solution we will have to dive deeper into the metaphysical underpinnings the problem is built upon. We will do so in the coming chapter. There, the fact that Bayesian representationalism predicts that Mary can have second-order knowledge will turn out to be relevant to a more comprehensive solution to the hard problem.

8.8 Summary

We have now finished developing Bayesian representationalism as a unified account of consciousness. Starting out from the phenomenological investigation of part one we assumed that states of consciousness are reflexive representational states that represent the world in terms of appearance properties. Chapter four argued that classical referentialist theories of mental representation have a hard time accounting for appearance properties.

In chapter five we introduced the paradigm of Bayesian cognitive science together with the objective Bayesian assumptions that underpin it. In the context of a

representationalist analysis of consciousness Bayesian cognitive science offers the intriguing perspective of a wholly representationalist analysis of mental phenomena.

Chapter six asked how it is that systems as described by Bayesian cognitive science may possess representational properties in the first place. Our answer was that the brain bears representational properties in virtue of its causal structure that mirrors the inferential structure implied by a probabilistic and (in virtue of an interventionist analysis of causality) causal structure of the represented environment.

Chapter seven engaged with a more thorough mapping of phenomenal properties onto representational ones. In particular we saw that low-level representations within the predictive hierarchy may explain how it is that conscious mental states represent the world in terms of seemingly primitive properties. If correct, this evinces the superiority of the inferentialist paradigm over classical referentialism.

Finally, the current chapter engaged with the remaining problem of explaining the apparent reflexive nature of consciousness and its special epistemological status. Central to our analysis was the contention that consciousness can be known through special introspective inferences. We still cannot explain consciousness in terms of brain states. To do this we will have to think about the mind-world relation as suggested by the paradigm of Bayesian cognitive science in more detail. This will be the central subject of part three.

Step Three:
Realism and Consciousness

9 The World From Within

Part three of the thesis will dive into the metaphysical implications of taking Bayesian cognitive science seriously. In particular, this chapter will argue that the Bayesian paradigm, particularly predictive processing, should make us rethink the mind-world relation as naturalists usually conceive of it. The following chapter will then return to the hard problem of consciousness in the light of our revised metaphysical views.

The default position arguably taken by most contemporary scientists and most contemporary philosophers is form of *metaphysical realism*, the view that the world as it really is is thoroughly mind-independent. Metaphysical realism is best illustrated by the metaphor of a *God's eye point of view*. If metaphysical realism is true, then there is a God's eye point of view on reality, meaning a point of view from which the world appears as it really is, independently of how it appears to contingent inhabitants like us.[1] Below we will take more time to spell out what this would entail.

Diametrically opposed to metaphysical realism is *relativism*, the view there is no true description of reality over and above those 'true for' a particular observer or group of observers. Relativism claims that reality is mind-dependent through and through and that the very same thought can be true from one perspective and false from another perspective, with no hope of ever reconciling the two.

Hilary Putnam's *internal realism* arises from the desire to reach some middle ground between relativism and metaphysical realism, avoiding pathologies of both extremes.[2] I will not argue against relativism here and assume that metaphysical realism of some form or other as a commonsensical starting point. But metaphysical realism faces the challenge of explaining how we could ever as much as refer to mind-independent reality. Even the metaphysical realist wants to hold that we sometimes refer to reality as it really is and maybe even sometimes get what we are saying and thinking right.

[1] Putnam 1977.
[2] Putnam 1981.

But metaphysical realism has trouble explaining the metasemantic relation that is supposed to hold between our minds and world as it is independently of our minds.

Against both extremes the internal realist holds that mind and external reality codetermine what is true. The idea here is that descriptions of reality are part of *conceptual schemes*, renderings relative to which particular observers parse reality. Truth is determined by fitting the external world against a conceptual scheme. An optimal fit against a conceptual scheme entails a true description of reality relative to this particular conceptual scheme. Truth then becomes nothing other than idealized justification within a conceptual scheme.

Internal realism differs from relativism in emphasising that truth may be relative to conceptual schemes, but *given* a conceptual scheme, what is true and what is false is completely determined by the world. To do this, the internal realist has to provide some criterion for the goodness of fit of the environment against a conceptual scheme that is itself not relative to any conceptual scheme. Thus the internal realist has to provide what Earl Conee has called an "unrelativised notion of fit"[3] or internal realism becomes indistinguishable from mere relativism.

This chapter will argue for a variant of internal realism. The nature of truth and the mind-world relation are infinitely deep issues, deeper even than the nature of consciousness. Thus my single chapter treatment will, by necessity, be wanting in some respects. The reader should keep in mind that the discussion is merely instrumental and its ultimate goal lies in laying the foundation for the discussion of the metaphysics of consciousness in the coming chapter.

Section one will explicate metaphysical realism. Section two will discuss Putnam's arguments against metaphysical realism, independently of the new ideas of Bayesian cognitive science. I will refute Putnam's famous brain in a vat argument against metaphysical realism but I will also hold that his model-theoretic argument is essentially valid. Then, in section three, we will see that it is hard to make sense of metaphysical realism if one assumes that our minds are best described as approximate Bayesian inference systems. Section four will introduce a variant of internal realism as inspired by Bayesian cognitive science which I call model-relative realism. Section five will engage with some objections to model-relative realism.

[3]Conee 1987, p. 90.

9.1 Explicating Metaphysical Realism

Metaphysical realism can be spelled out in a number of ways. I already mentioned the intuitive illustration that the metaphysical realists thinks that it is coherent to think about the world from God's point of view while the metaphysical anti-realist denies this. A more precise way of thinking about these matters however is in terms of *knowledge transcendence* or just *transcendence* for short. The metaphysical realists holds that truth may be transcendent, meaning unknowable, while the metaphysical anti-realists holds that there are no unknowable truths and thus that truth is *anti-transcendent*. Putnam characterizes anti-realism thusly:

> [L]et T_1 be an ideal theory, by our lights. Lifting restrictions on our actual all-too-finite powers, we can imagine T_1 to have every property except objective truth - which is left open – that we like. E.g., T_1 can be imagined complete, consistent, to predict correctly all observation sentences (as far as we can tell), to meet whatever "operational constraints" there are (if these are "fuzzy", let T_1 seem to clearly meet them), to be "beautiful", "simple", "plausible", etc. The supposition under consideration is that T_1 might be all this and still be (in reality) false.[4]

If we differentiate between *internal* and *external* criteria of truth, where internal criteria are those that Putnam mentions and all other similar criteria one may want to add, and external criteria are all other criteria on grounds of which one may judge a theory true, then the principle of knowability says that no amount of meeting internal criteria is sufficient to guarantee truth.

The transcendence of truth is closely connected to the idea of God's perspective. For if one accepts transcendence then this entails that things can be utterly different from how they seem even to observers in epistemically ideal situation. This could then only be known by an ideal observer that is, by definition, all-knowing.

Metaphysical realism holds that the world can be conceived as though from a point of view outside the world. In chapter five we saw that metaphysical realism is implicit in the standard way of thinking about the mind-world relation in

[4]Putnam 1977, p. 485.

Bayesian cognitive science. Let us revisit this point now that we have a better grasp on the nature of metaphysical realism.

For the present purposes we can bracket the fact that Bayesian cognitive science emphasises the approximate nature of real-life Bayesian inference and talk as if agents are actually Bayes-optimal. Generally, we saw that probability talk in Bayesian cognitive science is interpreted epistemically. When one says that an agent represents its environment as a probability distribution then this is standardly interpreted to mean that the agent represents different states the world might be in with different degrees of epistemic uncertainty. Probabilities just encode states of knowledge.

Under such an epistemic understanding of probabilities the Bayesian paradigm of cognition can make ready sense of even the most radically sceptical scenarios. For instance, the view that the true structure of reality is utterly different from how it appears to be *from all possible perspectives* seems coherent from this standpoint. An agent just has to assign a reasonably high probability to the proposition that reality has this or that transcendent structure. Thus orthodox Bayesian theorists, that is theorists who think of probabilities in purely epistemic terms, will implicitly accept the transcendence of truth.

This is why, in chapter five, I claimed that Bayesian cognitive science thinks of the mind-world relation in a metaphysically realist manner. Representational relations are, so to speak, assigned from the perspective of an outside observer (a God's eye point of view) correlating different internal states of the agent with states of the world. Perceptual inference is thought of as a process of figuring out which of the world states are more epistemically probable, given sensory evidence. But the true state of the world is a fact that is presupposed wholly independently of any agent's capacity to figure out what this state may be.

Remember that in chapter five I criticised the standard way of thinking about the mind-world relation on grounds of the fact that we had trouble making sense of its metasemantics. For all that a predictive mind 'cares about' regarding the external world is the mismatch between its impact on the sensory surface and the expected impact. This is why the semantics for predictive processing developed in chapter six held that conditions of satisfaction and thus content should be tied to the tendency to reduce prediction error. Under these circumstances there can't be differences in con-

tent that are not connected to differences in possible prediction errors. And thus there could be no difference in content between believing that one is a brain in a vat and that one is not if there is not also some difference in the expected sensory stimulation.

If however, there are many different states of the world that will be identical with respect to their impact on the sensory surface and their potential to cause prediction error then it is unintelligible how these differences should be reflected in representational content. The argument against metaphysical realism given below can be viewed as a generalization of this point: Metaphysical realism has trouble explaining how we could even as much as refer to a supposed world as it is from a God's-eye point of view. Only in this way can we make sense of the assertion that a representational state that is in alignment with available evidence and that performs its task in organic self-maintenance ideally could nonetheless be false. For the moment, let us, by way of introduction to anti-realist thought, review Putnam's famous arguments against metaphysical realism.

9.2 Putnam's Arguments

In this section we will discuss Hilary Putnam's two famous arguments against metaphysical realism. One is the *brain in a vat argument*, the other is the *model-theoretic argument*. Both arguments share a common structure. They are based on the assumption that *metasemantic naturalism* is correct, i.e. that there is some naturalistically acceptable way of explicating how representational states come to bear their content. Putnam usually formulates this point negatively by saying that we would not (and should not) accept views according to which reference is fixed by "noetic rays" or "magic"[5]. In essence Putnam doubts that, if there were a God's eye point of view, we could intelligibly refer to the world from such a point of view. And thus, and this is important, purported reference to the world from this point of view turn are shown to be semantically empty on metasemantic grounds. We never even had any metasemantic purport to such a God's eye point of view.

I should mention at this point that Putnam later revised his views on the failure of metaphysical realism. This change of mind was facilitated to a large extend by his

[5]Putnam 1981, p. 51.

later view that metasemantic naturalism may be evaded without thereby appealing to "magic" strictly speaking. In particular, Putnam claimed that representational properties may be supervenient on the scientifically accessible base reality without being reducible to it.[6] Discussing this view in detail would be beyond the scope of this dissertation. Here I merely want to point out that this point of view would entail that Bayesian cognitive science ultimately cannot fully explain our metasemantic capacities (because nothing non-semantic can). And, at any rate, it seems to me that Putnam's change of mind is insufficiently motivated. The following discussion is predicated on a form of naturalistic optimism that holds that our semantic capacities ultimately will be naturalistically explicable. Importantly, Putnam never changed his mind regarding the validity of the following conditional: If metasemantic naturalism is true then metaphysical realism is false.

The central point of this section will be that Putnam's anti-realist views have got less traction in the philosophical community than they deserved. In my estimation, bracketing sociological considerations, there are two reasons for this. First, there has been an overemphasis of the brain in a vat argument that Putnam arguably intended as a graphic illustration of his deeper point, but one that fails in an important way. Secondly, following interpreters such as Michael Devitt and David Lewis, there has been some confusion around Putnam's model-theoretic argument and specifically the so-called *just-more-theory-manoeuvre*. I will begin by discussing brains in vats and then continue to the model-theoretic discussion.

9.2.1 The Brain in a Vat Argument

Is it possible that you and all other sentient beings are really brains in vats floating in nutrient solution and being stimulated by a computer running a simulation of the actual world? At a first glance the answer obviously seems to be 'yes'. But the brain in a vat argument attempts to show that this cannot be the case if metasemantic naturalism is true.

The idea is that, were we brains in vats, then we could not refer to vats, computer simulations, and nutrient solutions. For it seems that envatted me does not bear a proper naturalistically acceptable relation to actual brains and actual computers

[6] Putnam 1993; Putnam 1994.

and so on. At best, envatted me can thus refer to *simulated* brains and *simulated* computers. But this will not suffice to properly represent the sceptical thought that I may be an *actual* brain in an *actual* vat. Thus it appears that what seems like a possible thought representing a possible scenario, namely that I may be a brain in a vat, is no such thought at all. For if it were true it would be devoid of meaning.[7]

Button has offered the following ingenious reconstruction of Putnam's reasoning (where 'BIV' is short-hand for 'brain in a vat'):

(1) The BIV's word 'brain' does not refer to brains.

(2) My word 'brain' refers to brains.

(3) I am not a BIV.[8][9]

Plausibly, if transcendence holds, then the brain in a vat scenario should be a genuine possibility. As it is hard to see how the brain in a vat could refer to the real brains, i.e. it is hard to see how to construct a reference relation in this case that conforms to metasemantic naturalism, it seems that the scenario is not a genuine possibility at all. This, Putnam concludes, should make us see that the anti-transcendence of truth is more plausible and metaphysical realism is false. The attempt to imagine a certain world from God's point of view has failed for principled reasons. Importantly, the argument hinges on metasemantic naturalism. We can make this explicit by giving the following more elaborate reconstruction:

(1.1) There is no magic. (Naturalism)

(1.2) The only way for the BIV's word 'brain' to refer to brains would be through magic.

(1) The BIV's word 'brain' does not refer to brains. (from 1.1 and 1.2)

(2) My word 'brain' refers to brains.

(3) I am not a BIV. (from 1 and 2)

[7] Putnam 1981.

[8] Button 2013, p. 118.

[9] We can also formulate Button's argument in more usual modal terms by using as the first premise "If I were a BIV then my word 'brain' would not refer to brains." It would then follow that it is not only factually true that I am not a BIV, but it would follow that it is not even possible that I am one.

Now let me disclose my biases. I find the view that we can rule out the possibility that we are brains in vats by philosophical *a priori* reasoning incredible. So incredible indeed that if the denial of metaphysical realism were tied to the view that brains in vats are impossible I would tend to discard metaphysically anti-realist views on those very grounds. So it is convenient that I think that are good reasons to reject the argument.

To see where Putnam goes wrong it is instructive to contemplate the first matrix movie. Protagonist Neo realizes that his world is actually a computer simulated reality and that he is really a brain in a vat. He discovers this first by a series of strange occurrences and then by actually being enabled to leave the computer simulation. The fact that all he is experiencing outside the simulation may be just a more elaborate trick of the simulators is besides the point. The important lesson is that there is something that would constitute compelling empirical evidence that one is or was living in a simulated environment.

We can make this a bit more explicit by drawing on a distinction between different ways in which the true description of reality may transcend our epistemic capacities. First, a truth may be *investigation transcendent*, meaning that nothing one may do would enable one to see that it is in fact true. Obviously, from Neo's perspective, the fact that he is envatted is investigation transcendent in this sense. If the relevant simulation is any good then nothing he does will allow him to uncover his predicament.

On the other hand, a truth may be *recognition transcendent* in the sense that nothing one could possibly observe would allow one to conclude that it is in fact true. For instance, it may turn out that Goldbach's conjuncture, the proposition that every number greater than two is the sum of two primes, is recognition transcendent, that is if it turns out the this is an undecidable proposition.[10] Facts are investigation if nothing one could *do* would suffice to provide one with evidence either way. Facts are recognition transcendent if nothing could even be counted as evidence either way. Importantly, the fact that one is a brain in a vat is *not* recognition transcendent, as is evinced by the fact that Neo is able to uncover his predicament.[11]

Representing an investigation transcendent but not recognition transcendent fact is in conflict with metasemantic naturalism only if we presuppose an overly verificationist

[10]In this case an anti-realist would of course deny that there is a fact of the matter here.
[11]I adopt the distinction between investigation and recognition transcendence from Tennant 1997.

metasemantics, namely one where grasping a proposition entails knowing how to get evidence for this proposition. But this is quite a restrictive view. One may also conceive a perfectly fine naturalist metasemantics that holds that representing some state of affairs merely involves recognizing evidence that this state obtains as such, if such evidence is provided. Specifically, premise 1.2 fails as soon as we emphasise the distinction between the two kinds of transcendence.

Putnam realizes that this is a possible response to his argument and he even acknowledges that it invalidates the brain in a vat argument. While he does explicitly say this, his tacit view seems to be that we can construct a stronger version of the thought experiment where it is assumed to be strictly impossible to ever discover one's predicament and thus where one's envattedness is properly recognition-transcendent.[12] I agree that in this case the brain in a vat argument may work after all, however it has to be questioned what exactly we are asked to imagine here. We must be dealing with a world where it is a *metaphysical necessity* that a certain computer simulation runs flawlessly. This seems to strain our (at least my) imaginative capacities and makes the whole argument much less straightforward. Either way, I will not build my case against metaphysical realism on this argument.

Let us see briefly how we can make sense of the brain in a vat scenario in the context of the predictive mind. The distinction between two notions of knowledge-transcendence makes it straightforward to see what a predictive mind would have to do in order to represent that it is envatted. Generally, a simulated reality will differ from a non-simulated reality in causal structure. While expected sensory input may be roughly the same (i.e. model evidence will be equal) because we assume that the simulation is running smoothly, the reactions to various kinds of incoming evidence will differ (i.e. priors and likelihoods will not be equal). A brain that holds itself to be envatted may infer from certain seemingly impossible observations that there is something wrong with the computer at base reality while a brain that holds itself to be an inhabitant of base reality may infer from such observations that somethings is wrong with it's perceptual apparatus. On the other hand, as we should expect, it makes no sense to suppose that a predictive mind represents itself to be envatted in a perfected brain in a vat scenario, because nothing could count as evidence for or

[12]Putnam 1981, p. 131, footnote 3.

against such a thesis. Both hypotheses would be associated with equal probabilistic models of the world and so they would count as the same hypothesis.

Now that we have seen that Putnam's most famous argument against metaphysical realism fails, let us turn to his second one.

9.2.2 The Model-theoretic Argument

Before diving into the weeds of the model-theoretic argument let me start with an exegetical remark on Putnam's work. In my estimation there has been an overemphasis on the brain in a vat argument over model-theoretic considerations both regarding their relative argumentative strength and regarding their role within Putnam's thought. In fact, it is reasonable to hold that the brain in a vat argument is supposed to be a mere colorful illustration of the facts the model-theoretic argument is intended to illuminate in abstraction. When Putnam discusses brains in vats, both in *Realism and Reason* and in *Reason, Truth and History*, model-theoretic considerations still take center stage. And, as we will see in a moment, the model-theoretic argument is built on the same naturalist considerations about reference as is the brains in vats argument.[13]

While for some time the model-theoretic argument has been dismissed it as a *non sequitur* it has recently reentered discussion.[14] The argument attempts to show that the conjunction of transcendence and metasemantic naturalism yield absurdity. I will first introduce, in very broad strokes, the model-theoretic results Putnam bases his case on. I will then give Putnam's original argument and discuss criticism raised by Devitt and Lewis and finally, I will show why their criticisms do not pay sufficient due to metasemantic naturalism.

Model-theory is the study of the relation of formal languages[15] and models. Models are structures described by formal languages. They are no related to the generative models of Bayesian cognitive science. *Structures* are set-theoretic constructions containing sets of objects together with properties, relations and functions defined over these objects. For our purposes anything you like can be a structure, from

[13]Putnam 1977; Putnam 1981.
[14]Douven 1999; Button 2013; Haukioja 2017.
[15]As second-order languages don't have expressive power that cannot in principle be captured in a first-order language (Dalen 2008, chapter 5) we will here solely discuss first-order languages.

cars and their colors, the natural numbers and the greater-than ordering relation to events ordered by causal and temporal relations.

Model-theory studies which sets of sentences can be said to capture the properties of which structures. If, for some set of sentences of a particular language, there is an interpretation function that maps the non-logical expressions of the language onto the elements of the structure such that all sentences come out true, then we say that the structure is a model for that particular set of sentences. In more prosaic language, model-theory studies which sentences can possibly be interpreted to apply to which domain of objects, properties, relations and functions. In the context of model-theory we call a *theory* a set of sentences that is closed under deduction, i.e. every logical consequence of the sentences is already part of the set. An example of an intuitive model-theoretic result is that the natural numbers turn out to be suitable model for Peano arithmetic.[16] A more topical example may be that a particular environment and its causal structure may serve as a model for representations forming a predictive hierarchy, considered as descriptions of a causal structure.

The model-theoretic result Putnam bases his argument on two model-theoretic results. Gödel's *completeness theorem*, in its model-theoretic form, tells us that every consistent theory has a model. That is, if a theory is consistent then it is possible to find a structure that makes all sentences of the theory true. Further, the *Skolem-Löwenheim theorem* says that every theory that has a model has a model of every cardinality.[17] For instance, as we know that Peano arithmetic has a countable model (the natural numbers), this theorem tells us that it will also be possible to have an uncountable model of Peano arithmetic (sometimes called "non-standard numbers"[18]). Together, these theorems tell us that, for every consistent theory, we will be able to find a model of every cardinality. This is the result that Putnam requires in his argument.

The model-theoretic argument can be presented in various forms.[19] The kind of model-theoretic argument I want to formulate is a form of *reductio*. It attempts to show that the only way metaphysical realism could be true would be if reference

[16] A good introduction to central results of model-theory with an emphasis on philosophical model-theoretic arguments is to be found in Walsh and Button 2018.
[17] Boolos and Jeffrey 1980.
[18] Dalen 2008 p. 114.
[19] For a systematic exposition of its variants see Button 2013.

would work by magic, i.e. non-naturalistically. So imagine metaphysical realism were true. Then the following situation should be possible. From the viewpoint of all its inhabitants the world appears to be a certain way. In particular, the world is seemingly perfectly captured by a certain theory that we will call the *mortal's theory* of the world. It may be helpful to imagine that this is some kind of future perfected physics. As the mortal's theory is internally perfect it is consistent and meets all observational requirements. Thus it not only perfectly predicts all past observations of all observers but is also does so for all future events.

But appearances, the metaphysical realist believes, may always deceive. So let's imagine that the world *really* is wholly different from the way it is depicted by the mortal's theory. Really the world is described by a noumenal physics that we will call the *divine theory* that is a description from God's eye point of view that captures how the world is really like. We need not imagine that anyone occupies the God's point of view, i.e. we need not assume a God to make sense of the scenario. All we are imagining is that, though it does not appear that way, the divine theory really describes the world as it really is. Its terms really correspond to the elements of the physical deep structure of the world, say. Of course the divine theory will have to be observationally adequate (it is, after all, true), but we may suppose that it otherwise seems hopelessly complex and inelegant from a mortal's point of view.

We can now set our model-theoretic machinery into motion to show that the stipulated metaphysically realist scenario is bogus. The metaphysical realist has to explain how it comes that the sentences of the divine theory correctly refer to reality while the mortal's theory does not. For model-theory tells us that we can always map the terms of the mortal's theory onto the elements of the real world as described the divine theory such that all sentences in the mortal's theory come out true. This is a consequence of completeness and the Skolem-Löwenheim theorem. To see this, first conceive of an arbitrary model the mortal's theory of the same cardinality as the intended model of the divine theory. Then use this model to construct an equivalent model over the elements of the real world, i.e. the domain of the intended model of the divine theory by defining appropriate predicates and functions on the domain.[20]

[20]Putnam 1977, p. 485.

Now it seems that both the divine theory and the mortal's theory can be mapped onto the world such that they may serve as a description of it. And here is where the first inkling should arise that the metaphysical realist tacitly appeals to magical reference: The only difference between the divine and the mortal theory must be that the divine theory corresponds to the elements of reality by divine *fiat*, while the mortal's theory does not.

So far the argument may not be (and should not be) very convincing. For it seems that the metaphysical realist can easily respond by holding that there may still be perfectly naturalist explanation for the fact that the mortal's theory refers to the true constituents of reality while the divine theory does not. Take as an arbitrary example of a naturalist theory of reference the view that reference requires causal connections. Then the terms of the divine theory will be connected to the constituents of reality by the right kind of causal chain while the terms of the mortal's theory are not. Wouldn't this be a straightforward account of how reference to the world from a divine perspective is naturalistically acceptable?

Here is where things get interesting. Putnam attempts to provide a general template against this kind of response. He says:

> Notice that a "causal" theory of reference is not (would not be) of any help here, for how 'causes' can uniquely refer in [sic!] as much of a puzzle as how 'cat' can, on the metaphysically realist picture. (Putnam 1977, p. 486)

Bringing out the generality of Putnam's response, whatever naturalistically acceptable relation-term the metaphysical realist employs to claim that reference of the divine theory as opposed to the mortal's theory is naturalistically fixed itself has to refer. And it is precisely reference that we are trying to fix. This move has come to be known as the *just-more-theory manoeuvre*.[21] Appeals to a causal or otherwise naturalistic theory, Putnam says, appeals precisely to the thing that is in question, namely metasemantic access to description-independent elements of reality.

If this were indeed the right way to understand Putnam then the model-theoretic argument would be clearly confused. For the obvious reply here is that it is not the *term* 'causality' that is held to fix reference but *causality itself*. Therefore the metaphysical

[21] Taylor 1991.

realist does *not* presuppose that the relation term refers. This in fact is the response famously given by Lewis[22] and Devitt.[23] Devitt suggests that reference may be fixed by causation while Lewis holds that reference may be fixed by naturalness of interpretation. So Lewis' idea is roughly that there is an intuitive notion of naturalness under which it is natural to hold that a subway map refers to the subway system but unnatural to suppose that it refers to some arbitrary set of entities and their interrelations dispersed across the universe. The details are not so important here. All that matters is that these authors hold that there is a relation that binds representations to their referents, independently of how we refer to these relations themselves.

So understanding Putnam's just-more-theory manoeuvre in the way just suggested evidently will not save the model-theoretic argument. It involves a conflation between the relation fixing reference and the term used to pick out this relation. If causation or naturalness fix interpretation then there can be no 'reinterpretation'. But there are good reasons to think that Lewis and Devitt are in fact misinterpreting Putnam here. For so far it is unclear how Putnam's argument depends on metasemantic naturalism at all. Putnam repeatedly tells us that, were one to believe that reference is magical, then one should not be bothered by model-theoretic considerations. But it seems that if reference were constituted by magic the just-more-theory manoeuvre, as interpreted by Devitt and Lewis, would be applicable all the same: The sentence "Reference is fixed by magic" would be subject to reinterpretation in just the same way as "Reference is fixed by causality" would be. In fact, Lewis even mentions this point as one of the things he thinks are mysterious about Putnam's reasoning:

> Why is it a better way to achieve determinate reference if we get cat Nana into grasp with our noetic rays than if we hold her in our hands? [...] We know what Putnam says if we try to base determinate reference on natural causal connection: the theory of causal constraint on reference is just more theory, as subject as any theory to overabundant, conflicting intended interpretations. But why are supernatural constraints exempt from parallel treatment?[24]

[22] Lewis 1984.
[23] Devitt 1983.
[24] Lewis 1984, p. 233.

I want to suggest an alternative understanding of the just-more-theory manoeuvre.[25] It is this: If reference is supposed to itself be the kind of phenomenon that one can have true theories about, then our theories about reference should be empirically justifiable. But this demand is in conflict with transcendence.

Let me make this fully explicit. Assume that the metaphysical realists holds that the sentence "Reference is fixed by C" is true, where C denotes some naturalistically tractable relation like causation of naturalness of interpretation. She also holds that metasemantic naturalism is valid and thus that "Reference is fixed by C" is empirically justifiable. And thus "Reference is fixed by C" will have to be entailed by suitable background assumptions together with the right observational data.

Here is where the problems for the metaphysical realist re-arise and the just-more-theory manoeuvre shows its full force. For if "Reference is fixed by C" is entailed by some set of sentences (assuming they form a consistent theory) then they will be so entailed quite independently of whether reference is in fact fixed by C! That is of course because in a world where reference is not so fixed we will still be able to find a mapping onto the elements of the world such as to make the compound theory true. Whatever fixes reference 'from the outside' cannot do so in a way such that we could be justified in holding that reference is so fixed. This is why metaphysical realism is incompatible with metasemantic naturalism, which claims that truths about metasemantics are just as scientifically discoverable as any other kind of truth. In the words of Tim Button, claims about reference as the metaphysical realists construes them lack "empirical content".[26]

It seems to me that, understood this way, the just-more-theory manoeuvre is perfectly valid. Metaphysical realists either have to abandon their posts or they have to admit that their realism is a kind of unjustifiable belief rather than a thesis made plausible by the way the world shows itself. The former option seems to be more rational.

Certainly the reader has noticed the parallels between the model-theoretic argument and the brain in a vat argument. Both stipulate that metaphysical realism and metasemantic naturalism are true and derive absurd conclusions by showing that precisely the possibility of radical scepticism undercuts the possibility of naturalist metasemantics.

[25]Here I have been inspired by Button's *Limits of Realism* (Button 2013). Similar views have been expressed earlier in Douven 1999. Putnam comes closest to making this fully explicit in Putnam 1990.

[26]Button 2013, p. 33.

Importantly, the model-theoretic argument is more general than the brain in a vat argument and is therefore immune to the objection from recognition transcendence: By stipulation, the divine theory of reality is both recognition and investigation transcendent. Nothing could ever count as evidence for or against such a theory. As long as we stipulate that the divine theory of the world is both investigation transcendent and recognition transcendent the argument will be valid. As metaphysical realists admit such transcendence, the argument decisively refutes their view.

So far we have investigated metaphysical realism on its own terms and discovered there is serious tension between transcendence and metasemantic naturalism. We will now go on to discuss the issue in the context of Bayesian cognitive science and Bayesian inferentialism.[27]

9.3 Metaphysical Realism and Cognitive Science

In this section I want to investigate metaphysical realism from the standpoint of Bayesian cognitive science. I will argue that cognitive systems, particularly predictive processing systems,[28] do not represent an independently given reality in the sense that there are truth about reality independently of their capacity to know these truths. My argument will be divided into two main parts. First, I will show that metaphysical realism is incompatible with Bayesian cognitive science, viewed from an inferentialist perspective. Secondly, I will argue that the same holds true if we reason from more orthodox structuralist metasemantic grounds.

[27]There is a further possible argument against metaphysical realism that results from evolutionary biology. A number of philosophers like Karl Popper and Thomas Nagel have argued that evolutionary biology has trouble accounting for our epistemic capacities that enable us to know mind-independent reality (Popper and Eccles 1977; Nagel 2012). This point has recently been made rigorous by Donald Hoffman and his team using evolutionary game theory. In essence, orthodox evolutionary theory tells us that natural selection will not care for representations based on their contribution to fitness rather than their capturing truth. As fitness depends on truth together with the nature of evolutionary context, fitness will always win out over truth in the long run (Hoffman 2020, chapter 4). Hoffman concludes from this that we should embrace a form of local scepticism regarding the world that we normally perceive. Popper and Nagel conclude that evolutionary biology is unfit to account for our epistemic capacities. But a further and arguably more parsimonious reaction may be to hold that the notion of a fully mind- or organism-independent reality is nonsensical. We never possessed the epistemic capacities Popper and Nagel thought we do.

[28]This definitely holds true for what I called free energy users, but also for free energy minimizers if we interpret their representational properties realistically, an issue I have left open.

From an Bayesian inferentialist perspective the refutation of metaphysical realism is relatively straightforward. Bayesian inferentialism was tied to the view that the truth-conditions of representations should be linked to their propensity to increase or reduce prediction error. We saw, by considering the example of the Müller-Lyre illusion, this account runs into serious trouble without an appropriate idealization clause. Our conclusion was that Bayesian inferentialism should conceive of truth as an idealized long-term minimization of prediction error.

This entails that the content of the mental representations in Bayesian cognitive science cannot be recognition transcendent, for recognition transcendence would entail that nothing could count as evidence for the relevant content and thus the relevant content would be independent of prediction error dynamics. In the limit, the view entails that, provided a maximal amount of evidence, the representations of a predictive mind will inevitably converge towards what is true by metaphysical necessity. We will refine these views below.

Those who have been convinced that we should indeed think of representational content in the context of Bayesian cognitive science in terms of inferential role will be drawn to agree with this anti-realist assessment. However, if one holds that Bayesian inferentialism is a dubious doctrine, then one will be prone to hold that the fact it is incompatible with metaphysical realism is a decisive refutation of Bayesian inferentialism rather than the other way around.

In our discussion of referentialist metasemantics we argued that such accounts are thoroughly incapable of illuminating how it is that perceptual states come to represent appearance properties. As the conjecture that perceptual states represent appearance properties is phenomenologically based, our anti-referentialist argument ultimately hinges on phenomenological observations. If these observations now lead us to reject so deeply engrained a view as metaphysical realism, it would only be cogent to reject our phenomenological analysis instead of metaphysical realism. I will now argue that reverting to referentialist metasemantics will not help us avoid anti-realist conclusions. Thus the anti-realist conclusions can be defended even on more orthodox structuralist-referentialist grounds.

A number of authors have defended that we should conceive of the representations involved in Bayesian cognitive science in structural terms.[29] Structural representations are internal states that play certain functions in virtue of a certain isomorphism to a represented domain. Now it may seem that structural representations may help explain how we can refer to a mind-independent reality. But this view is confused. Structural representations are no more compatible with metaphysical realism than are inferential ones.

To see this we have to revisit the central tenants of structuralism. We saw in chapter four that the mere existence of an isomorphism is certainly not sufficient to explain representational content. In short, the problem is that isomorphisms are symmetric, non-normative and abundant while representations are asymmetric, normative and sparse. Just as in the case of inferentialism, in order for structuralism to make sense it has to embed isomorphisms in some kind of teleological context. Representations essentially serve/have certain functions and the representational content of a state will depend on its particular teleological properties.

But now it seems that, assuming that free energy principle can serve as an underlying framework for explaining mental processes, all functions of cognitive states will be ultimately tied to the minimization of free energy or prediction error.[30] A mental state will serve its function in terms of its disposition to reduce free energy in the long run. A mental state will fail to serve its function by failing to so reduce free energy. Given this, the truth conditions of structural representations in the predictive hierarchy will never outstrip the dispositions to impact prediction error: A representation will serve its function if it does, in the long run, reduce free energy, it will not if it doesn't. Thus true representations will be free energy reducing, false ones free energy increasing. So we end up in the same place as we did starting out from inferentialist grounds. Truth conditions are tied to idealized minimization of prediction error and do not make sense in abstraction from prediction error dynamics.

The upshot is that that Bayesian cognitive science conceives of cognition as a process of minimizing free energy or prediction error which is strictly determined by the influx of information from the environment. This entails that there cannot

[29] Hohwy 2014; Gładziejewski, Paweł and Miłkowski 2017; Kiefer and Hohwy 2019.

[30] Hohwy has provided an explicit account of how teleological properties may be explained in terms of the minimization of free energy in Hohwy 2020.

be representational content that is recognition transcendent strictly speaking. As a result, independently of how exactly we conceive of the metaphysics of representation, agents will represent the world as a mere probabilistic source of sensory input. The true state of the world will be a probability distribution, conditional on the generative model and a maximal amount of evidence. In the long run, the beliefs of agents are guaranteed to converge towards the truth.

This argument parallels Michael Dummett's famous *manifestation argument* against metaphysical realism. Dummett argued roughly that grasping meaning requires grasping truth conditions and further, that the grasp of truth conditions must be ultimately grounded in behavioral dispositions in language games. Under these assumptions it seems that is unintelligible how we could ever as much as speak about recognition transcendent facts.[31] Similarly, we have just argued that representations discussed in the context of Bayesian cognitive science are ultimately grounded in the minimization of free energy - their truth conditions are ultimately manifestable as impact on prediction error. As supposed recognition-transcendent facts are irrelevant to prediction error they drop right through our metasemantic nets.

Dan Zahavi has recently pointed out the tension between metaphysical realism and radical mind-dependence of the perceptible world implied by Bayesian cognitive science. A number of writers have insisted that Bayesian approaches in cognitive science imply that the world as we perceive it is partly dependent on our perceptual apparatus.[32] But the very same writers also typically hold that the world as described by natural science is the way it is, quite independently of our cognitive access to it. But, Zahavi asks, wouldn't the logical consequence of the mind-dependence of the perceptible world be that the brain processes that are supposed to explain our cognitive capacities be mind-dependent in the same way as other perceptible objects are? That is, isn't there an obvious tension between the view that the perceptible world is dependent on our cognitive capacities and the metaphysically realist assertion that the physical world is out there, quite independently of how it appears to us, but still knowable by us?[33]

[31] Dummett 1975. For an overview and discussion see Tennant 1997, chapter 6.
[32] Frith 2007; Metzinger 2009; Hohwy 2014. A recent example not mentioned by Zahavi is Parr, Pezzulo, and Friston 2022.
[33] Zahavi 2018.

The tensions between metaphysical realism and predictive processing are nicely expressed in what may be called the first comprehensive textbook of Bayesian cognitive science. Here the authors say:

> The results of inference are not necessarily accurate in any objective sense. [...] (i.e., the organisms belief may not actually correspond to reality) for at least two reasons. First, biological creatures operate under limited computational and energetic resources[. ...] The second reason optimality may be thought of as subjective is that organisms operate on the basis of a subject's generative model of how their observations are generated, which may not correspond to the real generative process that generates their observations.[34]

It is the second point, the model-relativity of inference, that is important here. The authors tell us that even under subjectively ideal circumstances an agent's inferences will not guarantee truth, that is correspondence to reality. This evidently requires a transcendent notion of truth and the "generative process", the actual source of sensory stimulation, is a kind of reality viewed from a God's eye point of view. The implicit commitments to metaphysical realism are not hard to spot.[35]

The problem Zahavi is pointing out is quite similar to that pointed out in our criticism of metaphysical realism. That is, in stating and thinking about their views, Bayesian cognitive scientists refer to the "generative process" as though *their own* mental capacities are exempt from the limitations their own theories seem to entail. That is, they hold the cognitive systems they are describing to be constituents of mind-independent reality rather than of reality as it is reflected in their own generative models.

[34] Parr, Pezzulo, and Friston 2022, p. 22. Emphasis from the original removed.

[35] In the context of Bayesian cognitive science the problems of metaphysical realism become most evident in the philosophical project that has been called *Markovian Monism*. The central idea here is to use the Free energy principle in combination with metaphysically realist assumptions to construct a novel metaphysics of the mind. Very broadly speaking the idea is to describe the totality of reality as a complex *true probability distribution* or generative process. This process can then be subdivided into subprocesses by drawing appropriate boundaries around systems (so-called "Markov blankets"). Systems that are stable will then minimize their own free energy, provided a fitting generative model. (Friston, Wiese, and Hobson 2020) But given these assumptions it is unclear how our own representational faculties may ever as much as relate us to the underlying generative process. All we will ever have metasemantic access to will be the environment as it is reflected in our own generative model.

Zahavi's solution is to deny metaphysical realism an adopt what he calls *transcendental idealism*, a view he adopts from Husserl's phenomenology. Put very briefly, the idea is that it is a mistake to draw an all too sharp distinction between the world as we perceive it and the world as it really is. Rather, the way the world is and the way it appears are conceived as two sides of the same coin. Appearance and reality are not, as the metaphysical realist would have it, metaphysically independent. Zahavi says:

> Husserl's idealism is not a reductive idealism. Husserl is not a phenomenalist that seeks to reduce the world to a complex of sensations. His opponent is [...] the objectivist, who claims that reality is absolute in the sense of being radically mind-independent. To deny the latter [...] is not to say that reality really exists in the mind, or that it is an intramental construction, but that reality is essentially manifestable, and therefore in principle available and accessible to consciousness.[36]

In the terminology introduced above we may say that Zahavi, following Husserl, opts for a kind of phenomenologically based internal realism where it is of the nature of facts to be appreciable by a suitably equipped mind but where facts still are not thereby freely constructable.

Zahavi's point is that Bayesian cognitive science, interpreted as the view that the Brain performs Bayesian inferences using internal representations (a view Zahavi calls "representationalism", a term I haven't adopted here for obvious reasons), cannot serve as a satisfying underpinning for a metaphysics of the mind. The reason for this, in his view, is simply that Bayesian cognitive science conceives of cognitive processes as independently existing entities realized in physical brain but cannot in turn explain our metasemantic access to those very entities. Instead, Zahavi turns to the tradition of phenomenology that tries to ground the metaphysics of the mind in the directly given appearance of the world. The following section will attempt to show a different path by making Bayesian cognitive science coherent with a form of internal realism, thus demonstrating that Bayesian cognitive science is not in conflict with a suitable claim of anti-transcendence. The Bayesian cognitive scientist need not be an "objectivist" in Zahavi's sense who "claims that reality is absolute".

[36] Zahavi 2018, p. 57.

9.4 Model-Relative Realism

So we have seen that metaphysical realism is both incompatible with Bayesian cognitive science and incompatible with Putnam's model-theoretic considerations. The alternative, as I have hinted at a couple of times, is a form of internal realism that holds that there is an essential connection between how the world is an how it appears to the subject, or, more to the point, that it is an essential feature of reality to be knowable in principle. In this section we will develop a form of internal realism that coheres with the principles of Bayesian cognitive science.

Before developing internal realism in the context of Bayesian cognitive science let us investigate Putnam's positive views in a little more detail. Once again, on Putnam's view it is not the case that there is, as he at one point calls it, "a ready made world"[37] independently of the mind. Rather, what is true and what is real is co-determined by mind and environment. Calling the structures agents or societies impose on their environments *conceptual schemes*, we can say that for Putnam, the truth is a description of reality as it results from fitting a conceptual scheme against the environment.

On Putnam's view truth is just identical to idealized rational justification within a conceptual scheme. Importantly, this view is different from relativism where what is true is wholly determined by the mind or, more specifically, a specific language community. If this were the case then nothing would stop us from making arbitrary propositions true by changing the way we talk. In his discussion of the notion of existence or non-existence of objects Putnam makes this point quite succinctly. In the context of asking whether sets of two objects are themselves further objects he says:

> Once we make clear how we are using "object" (or "exist"), the question "How many objects exist?" has an answer that is not at all a matter of "convention". That is why I say that this sort of example does not support cultural relativism. Of course, our concepts are culturally relative; but it does not follow that the truth or falsity of what we say using those concepts is simply "determined" by the culture. But the idea that there is an Archimedean point (or a use of "exist" inherent to the world itself)

[37] Putnam 1982.

> from which the question "How many objects *really* exist?" makes sense, is an illusion.[38]

Importantly, this kind of balancing act between metaphysical realism (where truth is independent of conceptual schemes) and relativism (where truth is *solely* dependent on conceptual schemes) requires, as we said above, an unrelativised notion of fit. That, *given* a conceptual scheme and an environment, what is true or not has to be strictly determined. If what is and what is not a good fit were itself only relative to a scheme then the whole approach would collapse into relativism.

In Putnam's most detailed outline of internal realism, namely his *Reason, Truth and History*, Putnam develops such a notion of unrelativised fit. The idea here is that what is and what is not ideally rationally acceptable will ultimately be entangled with our values and goals such as to result in a notion of "objectivity for us".[39] A number of authors have suspected that Putnam's position is ultimately unstable and is bound to collapse back into relativism.[40] But instead of engaging this issue in further detail I will use Putnam's approach as a starting point for my own form of internal realism.

Putnam's project starts from two background assumptions that obviously differ from the approach we have taken in this thesis. Before embarking on a constructive treatment of internal realism in the context of Bayesian cognitive science we should make those explicit. First, Putnam thinks of truth and representation as a mostly social phenomenon. Therefore the kind of relativism that threatens to inflict his project is a kind of *cultural* relativism. On the other hand, Bayesian cognitive science starts from individual cognitive systems as natural units of analysis. If Bayesian cognitive science entails that what is true is relative to a generative model, the relativism that looms here is *individualistic*, i.e. truth is threatened to become relative to a single cognitive system and its generative model or a class of cognitive systems sharing a generative model.

I should mention that there have been attempts to apply the free energy principle to social groups.[41] That this is possible should hardly be surprising. The conditions of application of the principle are strictly functionally outlined and it

[38] Putnam 1987, p. 71, emphasis in the original.
[39] Putnam 1981, p. 55.
[40] Steinhoff 1986; Conee 1987; Vlerick 2014.
[41] Kaufmann, Gupta, and Taylor 2021.

solely describes how a complex systems has to behave in order to respond to external stimuli adaptively. If one can identify the equivalent of sensory states, active states and a generative model, then nothing stands in the way of describing certain social dynamics as approximate Bayesian inference. That said, I will here continue to focus on individual cognitive systems.

A second disanalogy between Putnam's approach and the one that we are following is that Putnam conceives of representations as conceptually structured. However we saw above that Bayesian cognitive science conceives of mental representations as holistic entities that purport to reflect the holistic causal structure of the environment without being separable into distinct concepts.

Abstracting away from these disanalogies there is a deep resonance between Putnam's views and the approach I am advocating. I will call this approach *model-relative realism*. Where internal realism holds reality to be the result of fitting the environment against a conceptual scheme, model-relative realism holds that the true probabilistic structure of the world is gleamed from fitting maximal environmental stimulation against a particular generative model and adjust the parameters of the model. Parameters are of course those values within a model that can be adjusted as a response to input, i.e. by learning. Thus, in other words, the true state of the world is will be reflected by an idealized representation after learning from a maximal stream of evidence.

We can also capture this in the language of probability theory as the claim that the *true probability* of some proposition E with respect to a model M is given by a limit in time, denoted by t. Sensory information at t is s_t. The true probability yields:[42]

$$P_{true}(E) = \lim_{t \to \infty} P(E|s_t, M) \tag{9.1}$$

One may wonder why the truth about E is captured by a probability function rather than an ordinary truth value. The reason is that there can be no *a priori* guarantee that there will be a convergence of P to either 0 or 1 in the limit. In such conditions, the true probability will be given by whatever intermediate value $P_{true}(E)$ converges to. We would then be dealing with non-epistemic probabilities. Arguably quantum mechanical accounts of nature offer examples of such non-epistemic probabilities.

[42]The approach is inspired by Douven, Horsten, and Romeijn 2010.

While probabilities have been introduced as an epistemic notion, it should not come as a huge surprise that on metaphysically anti-realist accounts the epistemic and the non-epistemic may blur. We will refine this approach below but for the purposes of the present discussion it should be clear enough.

We may apply further restrictions. For instance, we may require that true probabilities require infinitely fine-grained generative models. Models that only have finitely many parameters that vary in the course of learning will converge to some point because they can integrate no further information from the environment. It seems intuitively more plausible to hold that true probabilities will be those arrived at given infinitely fine-grained models. Less formally this would mean that the true state of the world is that arrived at by learning from a maximal stream of evidence, given infinite memory capacities.

Model-relative realism comes with an obvious criterion of unrelativised fit, namely the minimization of free energy. What is true relative to a certain model can be determined by performing approximate Bayesian inference in the light of a generative model and sensory stimulation arriving from the environment. There is also no threat of relativism here: Given a generative model, determining what is true is wholly up to the environment that provides sensory stimulations.

An interesting corollary of model-relative realism is that it entails that truth is ultimately *pragmatically grounded*. What is true and what is false is not independent of what is useful to achieve certain ends. The free energy principle conceives of cognition as a tool for enabling the self-maintenance of an organism. Good inference guarantees the maximization of model evidence and thus self-maintenance via active inference. So the very same activity that generates convergence to what is true on this account also ensures pragmatic utility relative to the ends of the organism, i.e. survival. The notion of truth becomes an idealization of the notion of a pragmatically useful view on the world.

Another interesting consequence of this pragmatist point of view is that not all models are created equal. That is because not all models will be equally successful in maximizing model evidence (which is sometimes expressed as the minimization of "surprise") and thus at the self-maintenance of the systemic boundary. Therefore some models will perform better than others. This however does not

entail that there is something like an ultimate model that would deserve the status of "the way the world *really* is". In fact, given some sensory input, for every successful model there are bound to equally successful models of a different structure. It is thus a futile move for the metaphysical realist to try to define truth in terms of idealized fit against a "perfect model".

A useful way of thinking about the model-relativity of the truths about reality is to think about organisms that have a simpler mental life than we do. Obviously the way a polar bear or an earthworm perceive their environment is quite different from the way we do. In a certain sense, different creatures inhabit utterly different worlds. Drawing on the work of Jakob von Üexküll, Daniel Dennett describes these worlds as "Umwelten" that are structured by "affordances", i.e. loci of relevance to the particular organism.[43] From the perspective of Bayesian cognitive science the structure these "Umwelten" will be encoded in generative models evolved by natural selection to make the particular organism competitive within a particular ecological niche.

The problem of naturalist philosophers like Dennett is that, after beautifully illustrating the organism-dependence of optimal representations, he also seems to imply that our specifically human cognitive capacities are somehow exempt from parallel treatment. For on his view our representations of ultimate reality are not to be thought of as dependent on our own human affordances or the shape of our evolutionarily developed generative model! Were Dennett to fully embrace his own naturalism about reference and truth he would have to abandon his implicit metaphysically realist assumptions (as shown by the model-theoretic argument and our Bayesian considerations[44]).

The upshot of the discussion in this section was that, if we think of cognition as a natural process, then we have to discard our metaphysically realist presuppositions. The more explicit we make our assumptions about the nature of cognition the more evident this conclusion becomes. Bayesian cognitive science makes the conclusion crystal clear: If we conceive of the process that lies behind sensory stimulation as an independently existing reality (a "generative process" that cannot be mapped onto our generative model) then the metasemantic relation between our mental states and this transcendent reality becomes elusive. Only if we assume that a true description

[43] Dennett 2017, p. 78-79.

[44] In light of Dennett's pan-Darwinian philosophy the argument mentioned in footnote 27 is also relevant.

of reality *just is* the result of ideally fitting environmental stimulation against a particular generative model we get an intelligible account of our metasemantic capacities. I will now discuss some objections to model-relative realism.

9.5 Objections

In this section we will discuss objections to model-relative realism. First, we will discuss whether model-relative realism is incoherent because it appeals to properties of cognitive systems that are themselves given independently of any model. Secondly, I will show that model-relative realism can give a satisfying account of the relation of different models. Thirdly, I will discuss how model-relative realism avoids phenomenalism. Finally, I will discuss whether model-relative realism falls prey to Fitchian objections.

9.5.1 Are Brains Model-Relative?

Our criticism of metaphysical realism was roughly that naturalist theories of cognition entail that cognitive systems do not enjoy unmediated metasemantic access to mind-independent reality. Applying this insight to one's own standpoint, it becomes clear that reference to mind-independent reality is empty. But it seems that the metaphysical realist can level a very similar attack against the model-relative realist: Doesn't the very idea of dependence of reality on features of a cognitive system require that that there be such a system in the first place. In other words, doesn't the model-relative realist require that there is an organism, a brain or a cognitive system "out there" in mind-independent reality? If so, then any justification of model-relative realism based on the results of cognitive science would be incoherent.

To evaluate this charge we first have to be very clear about what model-relative realism is and is not saying. The central tenet of model-relative realism is a form of anti-transcendence. What we are requiring is that all truths, under ideal epistemic conditions, be knowable. But model-relative realism is *not* saying that the world is *ontologically* dependent on the existence of minds, that is that the world or the objects it contains could not exist were there no minds for them to appear to. This

would be a too radical form of idealism to be plausible and indeed it could, on grounds of circularity, not be justified on naturalist or scientific grounds.

So are facts about cognitive systems ideally knowable on model-relative realism? Certainly. For, according to Bayesian inferentialism, cognitive systems are individuated in terms of their causal structure alone. Now cognitive systems precisely represent their environment in terms of their causal structure in virtue of the interventionist analysis of causality. Cognitive properties are thus the kind of properties that can be inferred using approximate Bayesian inference in relation to a generative model. Therefore the mental properties described by Bayesian cognitive science are anti-transcendent. Furthermore they are perfectly real independently of whether anyone actually comes to know about them.

This, our objector continues, shows that cognitive systems are knowable *given the right generative model*. But it does not show that facts about cognitive systems are independent of any particular generative model. For instance, earthworms presumably have very simple mental lives. Even under ideal epistemic conditions, we can reasonably suppose earthworms will not be capable of discovering any interesting facts about human cognition to speak of. So cognitive systems only exist relative to human and comparable cognitive capacities!

Here our objector has got it right. However, the model-relativity of facts about cognitive systems is itself benign. Cognitive systems, we may say, are part of our human ontology, the human Umwelt, but they are not part of the ontology of earthworms. But they are there objectively as, given the right model, the fact that there are humans and human minds and so on is not "up the mind" but "up to the environment". There is no problem of ontological dependence of cognitive systems on the mind and thus also no threat of an implausible amount of idealism implicit in model-relative realism.

9.5.2 Intersubjectivity

A further apparent problem for model-relative realism is that it has trouble explaining how it is that two cognitive systems can represent the same state of affairs. For it seems that truth that agent A represents will be relative to A's generative model, while truths that agent B represents will be relative to B's generative model and the point of model-relative realism was precisely that there is no model-independent

set of truths over and above these. This seems to entail that there is no intelligible way in which agreement and disagreement between agents may be possible, because this would require intending the same matters of fact in the first place.

The problem is further complicated by the holism implicit in Bayesian inferentialism. It is a well known problem of inferentialist accounts that they have trouble avoiding a kind of semantic solipsism where every agent inhabits her own cognitive world. For we cannot say that two states of different agents bear the same content in virtue of inferential role, unless these agents share all inferential roles, or unless we draw some principled distinction between meaning-conferring and non-meaning conferring inferential roles.[45] As it is unclear how such a distinction may be drawn in the context of Bayesian cognitive science we seem to be facing a problem.

The outlines of a solution to this problem in the context of Bayesian cognitive science have been given by Alex Kiefer and Jakob Hohwy in their discussion of structural representations. Structuralists face very similar problems as inferentialists do. Here too the inferential hierarchy is conceived as representing a holistic inverted mirror image of the causal structure of the environment.

The solution, the authors propose, is to move from a notion of *sameness* of meaning to a notion of *similarity* of meaning.[46] We have, for instance, already introduced the cross-entropy as a model for the divergence between probability distributions. We can then similarly use the cross-entropy to compare distributions between agents and thus assess the similarity and difference between two system's world-models. This is an important result for the model-relative realist because it gives us the ability to compare models without thereby implying that there is a unified underlying reality all agents uniquely intend.

There is a further interesting but not insurmountable problem involved here. Two agents may, intuitively speaking, enjoy a very similar outlook on the structure of the world while differing in the structure of their internal model. For instance, the divergence in models may stem solely from the fact that one agent sees the world from one corner of the room and the other one sees the world from another corner and another angle. An ideal tool for comparing probabilistic model-structure

[45] Lepore and Fodor 1993.
[46] Kiefer and Hohwy 2018.

would have to correct for such distortions without recourse to an underlying absolute perspective. However, it is plausible that such a tool can be conceived.

The problem of intersubjectively shared content also relates to the problem of how to explain the conceptual structure or mental content. We usually think of mental content as conceptually structured, i.e. as being composed of concepts. We already remarked that it is unclear how concepts may be reflected in the holistic mental content of Bayesian cognitive science. Here I at least want to gesture towards how such an account may be put together.

It is quite plausible that many conceptual contents can be explicated using functional language. For instance, an electron is precisely the kind of entity that relates to its environment in the way an electron does and it is arguably metaphysically impossible for there to be an entity that behaves causally like an electron in every context but that is not an electron. As agents represent their environment in terms of causal structure it is plausible to hold that they can represent the fact that electrons exist in their environment just by representing their environment as possessing a particular causal structure, that is a structure that contains entities that behave like electrons. Given a functionalist analysis of the content (technically this will be an intension) of some concept it is plausible that one could translate some conceptually structured content ("the cat is on the mat") in terms of the model-talk of Bayesian cognitive science (i.e. "the world has a structure of this and that kind").[47]

How far can we go with such an analysis? It seems plausible that we could in principle explicate many physical, chemical, biological, social and psychological concepts in this way.[48] I don't have much of interest to say about mathematical concepts and so I will keep silent.[49] If it is correct that propositions of the mentioned kind can be functionally analyzed, such an analysis would result in a mapping from causal structures, represented by cognitive systems, to sentences couched in the relevant vocabularies that would be true in these causal structures. Given such a mapping, we could associate cognitive systems with probabilistic beliefs over conceptually

[47]To do this explicitly it it would be helpful to rely on David Chalmers' scrutability framework developed in *Constructing the World* (Chalmers 2012a). In fact one may try to link up Chalmer's inferentialism (Chalmers 2021) with Bayesian inferentialism where the latter is presupposed as a theory of first-tier content and the former is used as a theory of second-tier content.

[48]For instance, Chalmers 2012b argues that many concepts can be functionally analyzed in this way.

[49]I briefly comment on the issue in the appendix.

structured propositions.[50] We could then say that two agents may both represent a particular conceptual content by representing the relevant similar causal structure.

Here it is also natural to ask whether this kind of analysis conflicts with arguments for semantic externalism. For instance, Putnam famously argued that no amount of information about facts internal to a speaker (or thinker) will determine whether this speaker (or thinker) intends water by his words (or thoughts). For reference to water is constituted by the causal connection to *actual water*, a connection that is not internal to the agent. We could, for instance, imagine a twin earth where the role of water is played by some water-like liquid composed of XYZ instead of H_2O. Twin earth agent's, prior to the advent of modern chemistry, would arguably refer to XYZ where we refer to H_2O without differing from us in internal structure in relevant ways. Thus no internalist analysis of the application of the concept of water seems possible.[51]

Arguably this is not a huge problem. Representational content that is supervenient on the skull is called *narrow content*, representational content that is partly dependent on the environment is called *wide content*. A satisfying theory of narrow content requires that such content, in combination with the environment, determine wide content. Again by relying on functional roles it is possible to spell out the narrow content of the thought that there is water in the glass: "There is the stuff in the glass that plays the water-role in my local environment." The wide content will then be "There is H_2O in the glass." on earth and "There is XYZ in the glass." on twin earth. This proposal can be spelled out more elegantly using two-dimensional semantics,[52] a topic we will engage with briefly in the coming chapter. For the moment it shall suffice to have shown that there are no deep difficulties involved in mapping conceptually structured propositions onto the content predicted by Bayesian inferentialism. Narrow content will be determined by the inferential hierarchy. Wide content will be determined by the hierarchy together with the environment, or more accurately, together with a potential totality of evidence from that environment, as the metaphysical anti-realist will deny that there are facts about the environment that aren't discoverable.

[50] Igor Douven and Leon Horsten have argued for an anti-realist definition of truth that defines it as a convergence of degrees of beliefs of rational agents. (Douven, Horsten, and Romeijn 2010) This approach would fit perfectly with the account suggested here.

[51] Putnam 1975.

[52] Stalnaker 1999; Chalmers 2003b.

9.5.3 Phenomenalism

A misunderstanding that I also want to rule out explicitly is that model-relative realism could in any sense be described as a form of *phenomenalism*. I understand phenomenalism broadly as the view that the relevant phenomenal or sensory states that serve as evidence for higher-level concepts are in some sense not just epistemically but *metaphysically* more basic than higher-level facts. For instance, a phenomenalist may hold that a cat really is nothing else but a certain pattern, actual or potential, in sensory activity or sense data.

Model-relative realism is thoroughly couched in the language of probability theory. The central idea is that the environment can be fully characterized by a certain true probability distribution that sensory states are sampled from or generated by. True, we have said that the true probability distribution will be determined by an ideal limit of sensory evidence, but this should not trick us into conflating sensory states with what they are evidence for. For first, they are conceptually distinct. Secondly, the shape of the true probability distribution will be equally shaped by the generative model as it is shaped by the incoming sensory states. Intuitively speaking, the cognitive system represents the environment as *the source of* sensory states. Model-relative realism is not and does not imply phenomenalism.

9.5.4 Fitch's Paradox

One of the arguably most influential objections to views that hold that the world is knowable in principle has been formulated by Frederic Fitch. It starts out from the assumption that some truths are not in fact known. For instance, probably no-one ever comes to know how many hairs my beard consists of, even though there is some definite number b. Even the defender of anti-transcendence shouldn't deny *that*. Let B the proposition that I have b hairs in my beard. Let us also introduce the $K(X)$ which expresses that someone at some point in time knows, that X. Then we can conjoin the fact that I have b hairs in my beard together with the fact that no-one knows this: $B \wedge \neg K(B)$. This, it seems, is a properly unknowable fact. For if it were known, then it would be false and consequently could not be known

(because one can only properly *know* what is true).[53] It seems we have constructed a counterexample to anti-transcendence from innocent assumptions.

One may hope that the there is a way to avoid Fitch's paradox by appealing to the superficially non-conceptual nature of content involved in model-relative realism. But this is hopeless. For, given that knowledge of the number of hairs in my beard can be reformulated as a particular complex functional relation between an agent and my beard, then we can resolve the Fitch sentence as the fact that no agent bears this particular causal relation to my beard. Still, this will be an unknowable truth for oneself cannot bear the this very causal relation either without making $\neg K(B)$ come out false. Not believing $K(B)$ will entail associating all $K(B)$ scenarios with low probabilities. For the remainder of the discussion of the Fitch paradox I will assume that we have a sufficient grasp on how conceptually structured propositional contents are represented by agents.

My discussion of these issues will be superficial and I will only hint at a possible solution. An in depth discussion would deserve a chapter of its own. However, given the centrality of Fitch's paradox in the literature, it would be question-begging not to raise the issue at all. I will keep the discussion short by bracketing the once popular solution favoured by Eddington[54] and, in the eyes of many, refuted by Williamson[55] as well as the solution proposed by Tennant[56] and also effectively criticised by Williamson.[57]

The diagnosis I want to offer of where the problem resides, and another possible solution to the paradox, may be gleamed from thinking about semantics of logical conjunction. B is perfectly knowable and so is $\neg K(B)$. $B \wedge \neg K(B)$ is not. The standard semantics for $\wedge$ is that $A \wedge B$ is true if A is true and B is true. On model-relative realism this arguably entails that $A \wedge B$ is true if A and B could be inferred to be true under optimal epistemic conditions. And this leads to the Fitchian paradox: It cannot be the case that B and $\neg K(B)$ can both be inferred under optimal epistemic conditions.

The problem is evidently that, under the epistemically ideal conditions under which B can be known are conditions under which $\neg K(B)$ cannot be known and *vice versa*. But there is a simple fix here. Why not introduce *two* epistemically ideal

[53] Fitch 1963.
[54] Edgington 1985.
[55] Williamson 1987.
[56] Tennant 1997, p. 272-276.
[57] Williamson 2000, though this is disputed in Tennant 2001.

conditions? I am currently in conditions under which $\neg K(B)$ can be inferred (I have reasonably good evidence to hold that no-one has ever counted the hear in my beard). Let's call these *K-ideal* epistemic conditions. But if I take a week off, shave off my beard, and count the hairs in the sink, then I will be in a condition where B can be inferred. Let's call such conditions *B-ideal* epistemic conditions.

We can modify our above definition, where E is some *atomic* (propositions are called *atomic* if they do not contain logical junctions) proposition. To repeat, the true probability will be:

$$P_{true}(E) = \lim_{t \to \infty}(E|s_t, M)$$

But with the constraint that, first, E is atomic, and second, s_t in the limit entails that E-ideal conditions are met. This also ensures that E-ideality will be a knowable property, too.

This suggests the following straightforward solution to Fitch's paradox. $B \wedge \neg K(B)$ *is* knowable in the sense that the first conjunct is ideally knowable under B-ideal conditions while the second conjunct is knowable under K-ideal conditions. Or more generally, we can say that, on an anti-realist semantics that avoids the Fitch paradox, $A \wedge B$ is true if A can be inferred under A-ideal epistemic conditions and B can be inferred under B-ideal epistemic conditions. For most ordinary cases of conjunctions of course A-ideal and B-ideal epistemic conditions will overlap. But in cases like $B \wedge \neg K(B)$ they will not. By weakening the sense in which conjunctions are knowable we have effectively circumvented the paradox.[58]

An important challenge to this solution is that is requires that we can successfully separate epistemic from non-epistemic conditions. For otherwise all conjunctions of empirical propositions will be trivially true on the proposed semantics. For instance, let C be the proposition that my cup is on the table. Then, without any restrictions on what kinds of epistemic conditions are relevant, $\neg C \wedge C$ can be true. For under the condition that I look over and see my cup broken besides the table ($\neg C$-ideal condition) the first conjunct can be inferred. Under the condition that I look and the cup stands on the table ($\neg C$-ideal condition) the second conjunct can be inferred.

[58]This is essentially a rediscovery of Dummett's solution to the paradox given in Dummett 2001. Dummett merely holds that atomic propositions are knowable, and conjunctions are knowable only insofar as their conjuncts are.

And thus $\neg C \wedge C$ is true on our anti-realist semantics for the conjunction. The problem is evidently that we haven't properly differentiated epistemic from non-epistemic conditions. That is, the conditions under which one can come to know the conjunct must be conditions that pertain strictly to what one is thereby enabled to know. I have no explicit theory of how this task is to be fulfilled on offer. But I also see no reason to suppose that this problem may turn out to be devastating.

I have offered a potential solution to the issue of Fitch's paradox and thus have hopefully shown that the objection, while interesting and important, is far from devastating for defenders if knowability.

9.6 Summary and Discussion

This chapter dived into the weeds of the metaphysical implications of Bayesian cognitive science as we framed it in the previous chapters. The result of our analysis was that predictive processing, taken seriously as an overarching principle for the study of the mind, strongly suggests that metaphysical realism is spurious. In a nutshell, representational properties are tied to their disposition to reduce prediction error and not by their capacity of mirroring some independently given reality. The resulting view was that the true state of the world cannot be meaningfully described independently of a generative model that underlies perceptual inference and that is determined by an organism's biological makeup.

There are a number of intuitive stances on this view. I am sure that some readers who agree with the conditional (if Bayesian cognitive science is accurate then metaphysical realism is false) will see the problem in the philosophical assumptions Bayesian cognitive science is built upon. It starts out assuming that cognition is best conceived by drawing a strict boundary between internal states ("the mind") and external states ("the world"), consisting of active states and sensory states, and then one ends up with the whole Cartesian apparatus of internal representations, divorced from reality. Isn't this just a naturalist way of stating the core principles of the metaphysical locked-in syndrome that is Cartesianism? Ultimately such concerns will, if they can be answered at all, be answered by future research in the mind sciences. My suspicion is that essential parts of the predictive mind will stand the test of time.

It may to some seem surprising to hold that an analysis of our best cognitive science may bear on so fundamental matters of metaphysics. But on reflection, the reverse is clearly more accurate: Studying the nature of cognition *must* tell us something on the nature of the cognized. For whatever true thoughts about reality we think, these thoughts are possible only in virtue of a certain cognitive structure. And thus philosophical analysis of cognitive science may not tell us about what exactly is true about the world, but it may well bear on the nature of possible thought. It has been remarked that Bayesian cognitive science, mediated by Helmholtz's inferentialists views on perception, "has its roots in Kant".[59] It should not surprise us that what has its roots in Kant may bear some Kantian fruits.

[59]Swanson 2016.

10 Ontological Indeterminacy

The *easy problems of consciousness* pertain to its functional properties and neural correlates. The *hard problem of consciousness* is the problem of explaining why any physico-functional states are associated with phenomenal consciousness in the first place.[1] The distinction is not supposed to reflect the intellectual difficulties involved in philosophy versus those involved in cognitive science. Rather, the easy problem is easy insofar as we have some rough conception of its solution. The hard problem is hard insofar as we seem to have no clue what a solution may look like.

This chapter will argue for a particular way of resolving the hard problem of consciousness that I call *ontological indeterminacy*. Ontological indeterminacy holds that particular views about the fundamental ontology of consciousness, like dualism, the view that consciousness is irreducibly mental, and physicalism, the view that consciousness is ultimately physical, are underdetermined by the facts. Consciousness is neither composed of physical stuff, nor is it made from an irreducibly mental substrate. There is no explanation of phenomenal facts in terms of some underlying ontology.

The argument will proceed as follows. Section one will discuss the interrelation between questions of metaphysical realism, metaphysical anti-realism and fundamental ontology. Section two will argue that conceivability arguments against physicalism implicitly assume a form of metaphysical realism and metasemantic anti-naturalism. Section three will argue that knowledge arguments succeed in showing that there are phenomenal facts that aren't entailed by any physical facts, but that this will not settle question about the fundamental ontology of the phenomenal. The argument concludes that, in fact, *nothing* settles questions about the fundamental ontology of the phenomenal and thus phenomenal consciousness is ontologically indeterminate. This relativizes the axiomatic naturalism of previous chapters to the view

[1] Chalmers 1995.

that the hard problem of consciousness does not require that we invoke any entities beyond a broadly scientific world view. But this does not mean that consciousness is just physical. In essence, what can be understood about consciousness can be understood in naturalist terms. But certain questions, I will argue, particularly questions about its fundamental ontology are ill posed.

10.1 Meta-Ontology

The chapter discusses question of the fundamental ontology of consciousness, i.e. the question of what kinds of ontological facts ultimately explain phenomenal facts, in the context of model-relative realism. Before diving into the weeds of anti-physicalist arguments I want to give some thought to the general question of the relation metaphysically anti-realist views and questions of ontology. We will be engaged with the meaning of ontological claims or *meta-ontology* for short.

A natural way of thinking about fundamental ontology tacitly assumes a form a metaphysical realism. When we ask, for instance, whether the world is all physical or whether there are abstract objects that are real but not physical then it is natural to conceive of this question as one about the furniture of the world, independently of any particular observer. When God looks at the world, does he see only physics or does he also see numbers, triangles and so on? Sure, even on this realist perspective particular ontologies may be preferred on grounds of their explanatory power, but it assumes no necessary connection between explanatory power and what ontological facts obtain. The assumption is that ontological theses explain *why* particular kinds of theories fit the facts well. For instance, a metaphysically realist physicalist will hold that physical explanations work so well *because* the world is ultimately composed of physical stuff.

The denial of metaphysical realism by itself does not entail any specific ontological views. But what it does entail is that there is a necessary connection between explanatory adequacy of some view and the ontology this view commits us to. For instance, on the metaphysical anti-realist's view, abstract objects over and above physical objects exist precisely if the view that there are such objects offers the best explanation for the facts of mathematical practice, or whatever explanada a theory of abstract objects may have. Here the correct ontological views are *entailed*

by the explanatorily ideal theories. Idealized explanatory success and ontological truths are taken to be two sides of the same coin.

In the context of the philosophy of mind *physicalism* is the view that systems have their mental properties in virtue of their physical properties. The mental properties we are concerned with here are of course phenomenal properties. So naturally, for the metaphysical anti-realist, physicalism will be true precisely if physicalism offers the best explanation for phenomenal facts (and all other mental facts). In a similar vein dualism will be true if some theory that posits irreducibly mental substances offers the best explanation for phenomenal facts, neutral monism will be true if some theory that posit some third substance offer the best such explanation, and so on.

Thus for some non-physicalist theory to win out over the physicalist it is not sufficient to show that physicalist explanations are insufficient to account for phenomenal facts. Rather, the non-physicalist also has to argue that a non-physicalist explanation actually gives us an explanatory advantage over the physicalist. For the anti-realist, the validity of a view is evinced by its having an explanatory edge over alternative theories.

Physicalism, dualism, neutral monism and so on all entail a commitment to *ontological determinacy* about the phenomenal. Ontological determinacy about some domain holds that there are determinate facts about the fundamental ontology of this domain. *Ontological indeterminacy* about the phenomenal mind then is the view that there are no determinate facts about the fundamental ontology of the phenomenal mind. *This* thesis will hold, by the metaphysical anti-realist's reckoning, if no ontological stance has a decisive explanatory edge over alternative theories. Thus ontological indeterminacy holds roughly if all theories of consciousness are equally bad. Note that ontological indeterminacy arguably is not a viable option for the metaphysical realist. For independently of any epistemic (i.e. explanatory) considerations, the metaphysical realist will hold that phenomenal facts are either explained by some other kind of facts, or these facts themselves are primitive constituents of reality. At least intuitively,[2] metaphysical realism entails ontological determinacy.

[2]Intuitively that is because I don't have any specific argument. Unlike Dummett I do not *define* anti-realism as the view that not every statement is either true or false, realism as the view that this is the case. Perhaps one may hold that God sees that some facts are indeterminate. My point is that it is hard to see a concrete reason for why one may hold this.

Ontological indeterminacy with regards to consciousness closely parallels *ontological pluralism*. Ontological pluralism with regards to the phenomenal mind holds that, in light of the fact that there are many equivalent ways of parsing the ontology of the conscious mind, all these ways are equally valid. In effect, where ontological indeterminacy holds that all ways of parsing reality with regards to consciousness are equally bad, ontological pluralism holds that a number of ways of parsing reality are equally good. This throws up the question which of these meta-ontological views the model-relative realist ought to adopt where ontological determinacy breaks down. Ought one to settle for pluralism or indeterminacy when determining factors are nowhere to be found?

A model-relative realist can accept that there are many different ways of parsing reality employed by different cognitive systems. However, this is not what is of issue here. The question we are asking is whether *one and the same agent* can hold mutually inconsistent views, as the ontological pluralist would have it, or whether inconsistencies ought to be eliminated, as the ontological indeterminist holds.[3] As the rejection of metaphysical realism by itself gives no licence to hold inconsistencies to be true the latter view must be correct. In fact the assumption of consistency is built into the very heart of objective Bayesianism. So, for the model-relative realist, where ontological determinacy breaks down ontological indeterminacy follows. The remainder of the chapter will argue that ontological determinacy about the phenomenal indeed does break down and thus that ontological indeterminacy is true.

10.2 Conceivability Arguments

As I already mentioned there are two popular kinds of arguments against physicalism. First, there are *conceivability arguments* that hold that, because changes in phenomenal properties without accompanying changes in physical properties are conceivable, phenomenal properties cannot be explained in terms of physical ones. Secondly, there are *knowledge arguments* that hold that, because no amount of physical knowledge let's one derive phenomenal facts *a priori*, phenomenal facts are non-physical. Knowledge arguments will be dealt with in the coming chapter. Here I

[3]This ambiguity is nicely brought out in the discussion of Putnam's views in Conee 1987 p. 85-86.

will argue that conceivability arguments can be rejected on grounds of metaphysical anti-realism. The following reconstruction is largely based on the work of Chalmers.[4]

There are a number of ways to motivate the conceivability claims that underlie the arguments and we have mentioned two of these before. For instance, it is intuitive to ask oneself whether the experiential quality of blue that I experience when I gaze into the sky is the same experiential quality that other people experience. This is arguably a corollary of the fact that it is perfectly conceivable that a perceptual apparatus that is functionally quite similar or even identical to mine is accompanied by an inverted spectrum. Even more strikingly, a philosophical zombie, a being that is physically identical to me but wholly unconscious, seems equally conceivable.

The relevant notion of conceivability relevant to issues of metaphysics is not one of daydreaming but one of *ideal conceivability*. In some sense I can conceive a triangle (in euclidean space) where the angles sum to 181°. But of course I cannot *really* conceive this if I give sufficient attention to the details of what I am imagining. So the relevant claims here are that inverted spectra and zombies are ideally conceivable. Luckily for the dualist it is plausible that these cases really are ideally conceivable, for in order to show that a conceivability intuition is not an instance of ideal conceivability one has to show that there is a logical contradiction involved. The claim that there is some non-apparent contradiction hidden in the idea of an inverted spectrum or a zombie seems implausible.

Why should we assume that there is a connection between what is conceivable and what can be explained in terms of what? The crux here is that there arguably is a connection between what is conceivable and what is metaphysically possible. For our purposes metaphysical possibilities are conceptual or epistemic possibilities, given their relevant information about the nature of the entities involved. And metaphysical possibility is closely connected to issues of explanation. This suggests the following reconstruction:

> (1) Phenomenal differences without physical differences are ideally conceivable.
>
> (2) If phenomenal differences without physical differences are ideally conceivable, then they are metaphysically possible.

[4] Chalmers 2009.

(3) If phenomenal differences without physical differences are metaphysically possible, then phenomenal facts cannot be explained in physical terms.

(4) Phenomenal facts cannot be fully explained in physical terms.

As I already said, (1) is quite plausible. At any rate, the burden of proof here lies on the physicalist. (3) is a truism about the nature of explanations. Premise (2) however is another matter. There have been a number of examples that seem to undercut the supposed connection between conceivability and metaphysical possibility. For instance, it is not metaphysically possible for water to consist of XYZ because, as science has shown us, it is of the nature of water to consist of H_2O. However it may be plausible to hold that, prior to the advent of modern chemistry, it was ideally conceivable that water could turn out to consist of XYZ! To see that this kind of response does not invalidate conceivability arguments we have to investigate the nature of conceivability in more detail.

10.2.1 Two-Dimensional Formulation

To straighten out the argument we need to differentiate two different notions of conceivability and correlated notions of possibility. Something is *epistemically conceivable* iff it is conceivable, given our epistemic state. Similarly, something is *epistemically possible* if it is possible, given our epistemic state. It seems clear that inferences from epistemic conceivability to epistemic possibility are benign. For instance, prior to the birth of chemistry from the womb of alchemy, it was epistemically conceivable that water was made of water spirits and it was also epistemically possible that water was made from water spirits.

On the other hand, we define *metaphysical conceivability* and *metaphysical possibility* as conceivability and possibility, given the metaphysical natures of things. As it is the metaphysical nature of water to be composed of H_2O, it is neither metaphysically conceivable nor metaphysically possible for water to be composed of water spirits. Again, metaphysical conceivability may serve as a guide to metaphysical possibility. A central difference between epistemic and metaphysical conceivability is that, what is epistemically conceivable is easily accessible. What

is metaphysically conceivable may be epistemically challenging (though, for the metaphysical anti-realist, not epistemically transcendent).

We can now reformulate the conceivability argument in two ways. Either we substitute all mentions of conceivability and possibility by instances of epistemic or by instances of metaphysical conceivability and possibility. Now the epistemic reformulation of the conceivability argument is spurious. For premise (3), the inference from epistemic possibility to lack of explanation is not cogent. Just because *for all you know* it is possible to have X without having Y does not entail that X doesn't explain Y. But the metaphysical reformulation is equally spurious. For the conceivability intuition is only plausible insofar as we understand it in epistemic terms. For what do I know what is and isn't conceivable, given one knows the metaphysical nature both of the physical properties and the phenomenal properties involved? Figuring these out is precisely the purpose of the whole discussion and thus can hard be assumed as a premise.

To build a convincing argument we need a further assumption that has to do with the peculiar nature of phenomenal consciousness. The worries about our first pass at formulating the argument resulted from drawing an analogy between the metaphysical nature of water and the metaphysical nature of consciousness. Water has a hidden metaphysical nature that limits the degree to which our conceivability intuitions bear on the problem. But, on reflection, it turns out that consciousness is quite different from water in this regard. Specifically, it seems that our phenomenal concepts, like phenomenal blue, pick out precisely the way that this experience appears to us. No interesting appearance-reality distinction is to be drawn here, or in other words, the metaphysical nature of phenomenal consciousness and its surface appearance are one and the same thing! In this way, I may imagine a water-like substance composed of water spirits, but it turns out that what I was imagining wasn't really water at all. But it cannot turn out that when I imagined pain in my knee it really wasn't pain at all.

We will call the coinciding of appearance and metaphysical reality in the case of phenomenal properties the *revelation thesis*. In terms of our correlated versions of conceivability and possibility we can express the revelation thesis as the fact that in the case of phenomenal properties we can infer metaphysical possibility from mere epis-

temic possibility because in being conscious the metaphysical nature of consciousness is directly revealed to us. We then get the following version of the argument:

> (1) Phenomenal differences without physical differences are ideally epistemically conceivable.
>
> (2) If phenomenal differences without physical differences are ideally epistemically conceivable, then they are epistemically possible.
>
> (Revelation thesis) If phenomenal differences without physical differences are epistemically possible, then they are metaphysically possible.
>
> (3) If phenomenal differences without physical differences are metaphysically possible, then phenomenal facts cannot be explained in physical terms alone.
>
> (4) Phenomenal facts cannot be fully accounted for in physical terms.

So assuming the revelation thesis it is plausible that phenomenal consciousness cannot be physically explained.[5]

In light of our meta-ontological discussion above it is important to ask what the conceivability argument establishes for the metaphysical anti-realist, given we accept the revelation thesis. Focusing, for the moment, on physicalism and dualism as the only contenders as forms of ontological determinacy, does the conceivability argument licence the assumption of a separate irreducibly mental domain on metaphysically anti-realist grounds? Arguably it does. For dualism does seem to have an explanatory advantage over physicalism here: Given that there are ectoplasmatic constituents of reality, say, that, by stipulation, necessitate the phenomenal facts, the dualist can explain the phenomenal facts and has do quarrels with the conceivability of zombies. So if the conceivability argument would go through it would support some kind of non-physicalist ontological determinacy, regardless of whether metaphysical realism is correct. But as we will see in the next section, the conceivability argument fails.

10.2.2 Against the Revelation Thesis

To see that the conceivability argument is predicated on a form of metaphysical realism note that the phenomenal differences the thought experiments ask us to

[5] ibid.

imagine are recognition transcendent. *Nothing* would count as empirical evidence for the supposition that your neighbour actually is a philosophical zombie or has an inverted color spectrum. That is because every conceivable experiment we would subject our neighbour to, and every observation we could possibly make, would just reveal more functional properties. But the idea is precisely that phenomenal properties can vary independently of these functional features!

Even if we stipulate that we have a science fiction devise that we may call a *phenoscope* that connects two brains and causes experiences in the brain of the experimenter that, as far as possible, mirrors activity in the brain of the experimental subject this will not help solving our predicament. For how could we know whether the phenoscope is working properly and gives you the same experience as the subject? Imagine there is little knob at the side of the phenoscope that, when twisted, slowly inverts the color spectrum. How could the experimenter ever know what the correct position of the knob is such as to experience color exactly the way the subject does? The metaphysical realist may insist that there is an answer here, but if there is it seems to be recognition transcendent. Thus the metaphysical anti-realist will in turn insist that it is unclear how we could refer to the purported phenomenal differences.

This establishes that the phenomenal differences involved in the conceivability arguments are recognition transcendent *from the third person perspective*. But here the dualist may hold that this is due to the peculiar nature of phenomenal properties: They aren't publicly observable. But still phenomenal differences without physical differences would be knowable *from the first person perspective*. *I* know that I am not a zombie and *I* know that my color spectrum is currently non-inverted.[6]

Physicalists like Dennett have long pointed out that this is confused. The epistemological problems here are not one of first or third person perspectives but are of principled nature. For of course I could *not* realise that my own color spectrum was inverted, if my memory was similarly manipulated![7] Recognizing a color inversion requires comparing color qualities now to color qualities of past experience and so an inversion, when it is total, is fully recognition transcendent independently of one's particular experiential perspective.

[6]This kind of phenomenal epistemology is defended in Chalmers 2003a.
[7]Dennett 1988. The difficulties are also argued against Ayer in Putnam 1990, chapter 3.

A dualist may grant this possibility hold that we can pick out a particular phenomenal feel indexically ("Blue feels like *this*"). But knowing such indexicals is insufficient for the phenomenal facts that are supposed to drive the conceivability arguments are phenomenal *differences*. Knowing that "Blue feels like *this*" is true will not get the conceivability argument off the ground.

Note that it would be question begging by the metaphysical realist to resolve these questions by assuming that we can still imagine someone how has the intuitive power to directly know the phenomenal properties of other people. For the question at issue is precisely whether there is anything to know in the first place. To see where the error lies, compare the case to one where we all agree that there is no fact of the matter, like which direction, cosmically speaking, is north. Here one may reason that *we* might be unable to know which direction is cosmic north, but we can still imagine someone with an intuitive ability to know which point at the night sky is the universal north-pole. But this would be obviously flawed reasoning! We are sceptical about whether there are facts of the matter regarding which direction is cosmic north, precisely because nothing could possibly count as evidence for or against the view that some particular direction is the one we are looking for and so it is unclear which position such news would have in the inferential network of our mental states. In the same way, we should be sceptical about the idea that someone with telepathic skills could help us resolve the issue at hand.

If we accept the relevant conceivability arguments then we are committed to embrace a form of transcendence. If, on the other hand, anti-transcendence holds, as I claim to have shown in the previous chapter, then there must be something wrong with the conceivability arguments. Specifically, the arguments must be in tacit conflict with naturalistic metasemantics. And indeed there is such a conflict as I will attempt to show now.

Why does it seem so intuitive to hold that inverted spectra, say, are conceivable? Intuitively, it seems, we can intend a specific phenomenal property by some kind of "feels-the-same-relation". One, as it were, mentally points at a specific experience and thinks "... feels like *this*". On this view I can think of blue-experiences just by singling out all experiences that feel in *this* unanalyzable way.

However, from the standpoint of the present discussion it should be clear that the occurrence of the "feels-the-same-relation" is recognition-transcendent. As we just saw, nothing would count as evidence that a particular experience bears this relation to an experience in another subject or even to myself in the past or future, at least if we think of "feeling the same" in abstraction from the functional feature of the relevant experience which include one's dispositions to judge as same and different.

To make this point more explicit we can differentiate between two kinds of phenomenal properties. Let us say that *intrinsic phenomenal properties* are phenomenal properties that may vary independently of an agent's functional features. The property shared by a tomato experience in me and a experience of fresh grass in my inverted twin is an intrinsic phenomenal property. We could only refer to these properties if the feels-the-same relation were an additional ingredient in the metaphysical structure of reality, which the metaphysical anti-realist will deny. Intrinsic phenomenal properties are often called "qualia" in the literature, a term I will avoid due to the inexactitudes connected with it.[8]

Intrinsic phenomenal properties can be differentiated from *relational phenomenal properties* which I define as phenomenal properties that are tied to the functional roles of the mental states that bear them. These functional roles will include dispositions to recognize mental states as phenomenally equal, similar or different as well as maybe some other cognitive roles. It is not necessary to be wholly exact here and there may be a number of ways of picking out relational phenomenal properties. The question whether relational phenomenal properties *just are* functional properties will be left open until the coming section. Bayesian representationalism explains relational phenomenal properties in terms of low-level representations that represent appearance properties. These are, as we require here, functionally individuated and an inversion of low-level representations would necessitate functional changes and it is inconceivable for things to be any other way.

Let us now return to our two-dimensional reconstruction of the conceivability argument. We can either interpret phenomenal properties intrinsically or relationally. If we interpret phenomenal properties relationally then the first premise is false. Changes in relational phenomenal properties necessitate functional changes.

[8]For a nice survey of the confusion around this concept, see Stoljar 2004.

The only way to defend the conceivability argument is to interpret phenomenal properties intrinsically. But now it is the revelation thesis that turns out false. For the relevant revelation thesis holds that the metaphysical nature of intrinsic phenomenal properties is directly revealed to us. However, our metasemantic analysis has shown us that *we only* seem *to be capable of referring to intrinsic phenomenal properties*. Once we assume that metasemantics is naturalistically constrained and the feels-the-same relation is a spurious ground for reference we should admit that the metaphysical nature of intrinsic phenomenal properties is empty.[9]

The result we are left with is that the dualist was correct in claiming that we can, not merely superficially, imagine all kinds of intrinsic phenomenal changes without corresponding physical ones. But this does not show their metaphysical possibility because the kinds of properties we are imagining to change likely to not exist! Their metaphysical nature is empty. We will have to ask whether this is a form of illusionism in due time.

The diagnosis I am suggesting here is that the dualist, or whoever else is prone to wield a conceivability argument against the physicalist, tacitly assumes that we can think about phenomenal properties from God's point of view. Thinking about it in this way it seems unproblematic to imagine the experiential perspectives of other minds by putting ourselves in their shoes. It thus seems that there is an unambiguous fact regarding whether some experience of you and some experience of me feel the same, independently of their functional role in cognition. In light of our metasemantic naturalism it turns out that it is hard to see how such thoughts could be meaningful.

I have three more points to make before I bring the discussion of conceivability arguments to a close. First, it is instructive to ask whether a devout dualist should be moved by my argument to drop the conceivability argument. For, in many of its forms, dualism comes with a commitment to anti-naturalism at any rate. If you believe that your brain is imbued with consciousness by its relation to an eternal soul,

[9]This also suggests an interesting approach to what has been called the *meta-problem* (Chalmers 2018) of consciousness. The meta-problem is the problem of explaining why there seems to be a problem of consciousness at all. The strategy followed here is along the lines of Wittgenstein's famous remarks that, on the first glance we seem to understand the notion of what time it is on the sun. But, looking closer, we realize that we are dealing with a case of *apparent* semantic purport (Wittgenstein 2016, paragraph 350). The present discussion indicates that at least some aspects of the problem of consciousness arise because the semantic purport to concepts that create the problem of consciousness is apparent purport only.

then you arguably will not be compelled by arguments that invoke metasemantic naturalism as a premise. On the other hand, many modern dualists hope to explain consciousness as a minor addition to our theories of physical reality. For instance Chalmers, who even calls himself a "naturalistic dualist"[10], holds that consciousness may be a irreducible aspect of information processing structures that, in some way to be further elaborated, fits neatly into what we already know about the world.

If my argument is correct then a dualist should be moved by it only insofar as she wants to hold that representational states can ultimately be explained by the tools of cognitive science. If one believes that representational states are the result of complex organism-environment relations together perhaps with internal computational states, then one cannot add "... but *also* it is metaphysically possible to have phenomenal difference without physical ones." For *if* there were such differences we need an explanation of how we are able to think about the relevant properties in terms of our naturalistic metasemantics. And this, we argued in this and the previous chapter, cannot be done.

Secondly, the case of intrinsic phenomenal properties may serve as a model case for how to think about the limits anti-realism imposes on the space of the possible. According to two-dimensional semantics (of the form favoured by Chalmers) the meanings of concepts may be captured as mappings from possible worlds to extensions (technically called *intensions*). In order to capture the fine-grained differentiation pointed out by Kripke and others, like the fact that water possesses a metaphysical nature distinct from its surface appearance,[11] it makes sense to think these mappings two-dimensional. So here an intension of a concept is conceived as a mapping from a *world of utterance* and a *world of evaluation* to an extension. For instance, in the case of 'water', the world of utterance is the actual world where water turned out to be H_2O. The extension of 'water', as uttered in our world, will be H_2O in every possible world.[12] In the following matrix @ is the actual world. Columns represents extensions across worlds of evaluation, lines represent extensions across worlds of utterance.

[10]Chalmers 1998.
[11]Kripke 1980.
[12]Chalmers 2006b.

water	@	Twin-earth
@	H_2O	H_2O
Twin-earth	XYZ	XYZ

In the context of metaphysical anti-realism we can understand the possible worlds we are dealing with as possible constructions thought and language allow for. These are, so to speak, scenarios that we can talk about on a superficial level without any guarantee that our words correspond to anything. I suggest calling these *notional worlds* or *notional possibilities*. For instance, both a world where there are intrinsic phenomenal properties and a world where there are just relational phenomenal properties are both notional worlds. Both are scenarios that, on a superficial level, we can think and talk about.

We can now think of the constraints that metaphysical anti-realism and metasemantic naturalism put on the space of meaningful thought as constraints on the space of notional worlds. In the case of water, it turned out that the metaphysically possible is more constrained than the epistemically possible. In the case of metaphysical anti-realism it turns out that the metaphysically possible is more constrained than the notionally possible. For instance, the notional possibility of an inverted spectrum turns out not to be a metaphysical possibility.

We can also express all this in tabular fashion. The assumption we make now is that the world of utterance (vertical) constrains our metasemantics. We begin by considering the Nonvert world, i.e. a world where metasemantics isn't naturalistically constrained and spectra are non-inverted. Then the concept of intrinsic red will pick out red experiences in all worlds. Further, from this world, inverted spectra are possible. And of course in an inverted world the concept of intrinsic red, as it picks out precisely one kind of experiential quality, will also pick out red experiences in an inverted world (see the bottom line). However, if we now consider that, in the actual world (@), there is no way of even as much as referring to putative intrinsic phenomenal properties it turns out that the metaphysical nature of these properties is empty (line of @). The way to express this is to say that our world turned out in such a way that certain notional possibilities, possibilities for which's construction our mental apparatus and language allowed for, did not turn out to be

proper metaphysical possibilities at all. The concept of intrinsic red, as used in the actual world, refers to nothing, not even in the non-inverted or the inverted world.

The diagonal of a matrix captures its extension in all notionally possible worlds. One may heuristically express the content of the diagonal in the case above as "whatever the concept 'water' refers to in this world". The horizontal line within a matrix captures extension in all metaphysically possible worlds. This is why the idea of the revelation thesis can be nicely expressed as the thesis that for phenomenal concepts horizontal and diagonal intensions align.[13] And indeed, if we take the anti-naturalists point of view for a moment and erase the actual world from the table below we see immediately that this is indeed the case. Ignoring the possibility that one may err about metasemantics makes the revelation thesis seem plausible.

Once we take into account constraints on metasemantics we see that the revelation thesis is not true. For, from the perspective of the actual world, concepts that putatively pick out intrinsic phenomenal properties will not pick out anything at all - in any world! This is precisely the result of our previous discussion, namely that certain notional possibilities like spectrum inversion are no such possibilities at all.

intrinsic red	@	Nonvert	Invert
@	∅	∅	∅
Nonvert	red	red	red
Invert	red	red	red

The tabular notation discussed here does not really add anything to the arguments above. All I wanted to show is how we can think about constraints on metasemantics in analogy to familiar techniques from analytic metaphysics. In particular, discovering how a certain concept is metasemantically constrained is quite similar to discovering its metaphysical nature. In both cases one discovers that the meaning of one's words and thought are partly determined by factors extrinsic to one's mental and linguistic abilities.

Finally, my refutation of conceivability arguments does not entail a commitment to illusionism for at least two reasons. First, the revelation thesis may still be true for relational phenomenal properties. This would entail that, in a certain sense, consciousness really does have all the properties it appears to have. The philosophical problems arise

[13] Chalmers 2003a.

when one starts theorizing about experience, or, more precisely, when one starts assuming that there are facts about similarity of experience independently of functional similarities. Bayesian representationalism can capture the validity of the revelation thesis as it pertains to relational phenomenal properties in terms of the transparent inferential access, discussed at length in chapter eight. If relational phenomenal properties are representational properties then there is no reason to think that these representational properties aren't epistemically fully accessible in introspection.

The second reason why the elimination of intrinsic phenomenal properties does not entail illusionism is that this view is compatible with the assumption that there is irreducibly phenomenal knowledge, an issue we will discuss in the coming section. So far we have argued that the conceivability argument fails to threaten physicalism if one rejects metaphysical realism. Let us now see whether the knowledge argument fares any better.

10.3 Knowledge Arguments

The second reason for believing that consciousness poses a hard problem is the knowledge argument. It claims that, because no amount of physical knowledge allows one to derive phenomenal knowledge *a priori*, phenomenal facts are non-physical facts. As we have mentioned before, its most popular form is the thought experiment of Mary the neuroscientist which we will rely on for discussion here.

The knowledge argument does not presuppose the existence of intrinsic phenomenal properties. For if Mary gathers knowledge about relational phenomenal properties only, all the steps of the knowledge argument are left standing. A metaphysical anti-realist who is committed to physicalism has to give an independent account of how the knowledge argument can be avoided.[14]

If the rejection of intrinsic phenomenal properties does not entail that the knowledge argument is fallacious this also gives another clear sense in which the anti-realist response to the conceivability argument does not entail illusionism. A qualia freak may hold that, if the properties of consciousness are somehow tied to functional properties, then we have already lost the interesting part of consciousness from view. But this

[14]Crane 2003 makes the related point that the knowledge argument is not committed to qualia.

is not so: We still haven't given an account of Mary's supposed knowledge of non-physical facts. What remains of the concept of phenomenal properties is still strong enough to threaten physicalism about consciousness and thus is hardly uninteresting.

I will here assume that independent critiques of the knowledge argument are spurious. Famously, Lewis has argued that what Mary really acquires is a new skill rather than factual knowledge.[15] Dennett has said that the argument rests on an overestimation of our capacity to judge complex counterfactual scenarios *a priori*, for what do we know what a superhuman neuroscientist could and could not deduce by mere logic?[16] A committed physicalist may use these arguments to oppose the argument, however I think both are *prima facie* unconvincing and so the metaphysical anti-realist should be interested in how her anti-realism bears on the argument.

My argument will proceed in two main steps. First I will argue that the knowledge argument, rather than being an argument for a particular ontological point of view, should be viewed as an ontologically neutral argument for anti-objectivism, the view that certain facts can only be known if one inhabits a certain point of view. Secondly, I will argue that, rather than supporting physicalism, it entails that ontological questions about the ultimate nature of consciousness are underdetermined by the facts.

10.3.1 Anti-Objectivism

In this section we will investigate why one may be weary of the contention that the knowledge argument straightforwardly entails ontological consequences. As a number of commentators have pointed out, it seems that the knowledge argument can be leveled against the dualist as well as against the physicalist! For suppose Mary were given a comprehensive textbook on non-physical ectoplasm. Then she could learn everything there was to know about this strange mental substance. However it still seems true that this will not enable Mary to know what it feels like to see blue. And thus the knowledge argument can be leveled against the view that it supposedly establishes![17]

Jackson has replied that dualist Mary, as opposed to physicalist Mary, may at least know *some* phenomenal facts before leaving her room. For instance, while she may not know what seeing blue feels like, she may know that there is some experiential quality

[15] Lewis 1983.
[16] Dennett 1991.
[17] Churchland 1989; Crane 2003; Howell 2007.

that she will learn about.[18] But this will not do. Arguing against the physicalist, if there is just a single phenomenal fact that physicalist Mary can't infer *a priori* then physicalism is false. Similarly, if there is a single phenomenal fact that dualist Mary can't infer *a priori* then dualism must, by parity of reasoning, be false as well.[19]

Also, Jackson is mistaken to think that the knowledge argument entails that physicalist Mary could not know *that she will learn something* upon leaving her room. The argument is strictly speaking neutral on this question and should not push our intuitions either way. The only thing it seems to show is that Mary cannot know some specific experiential qualities beforehand.

As we saw above, Mary gathers first-order knowledge, knowledge about an experiential quality, upon leaving her room. But this does not commit us to say that she did not possess second-order knowledge, knowledge that she would learn something, all along. The knowledge argument seems to show that Mary could not have first-order knowledge but it is silent on second-order knowledge and this seems to be enough to trouble the physicalist. Thus, as far as the knowledge argument goes, dualist Mary has no edge on physicalist Mary.

If the knowledge argument does not show that consciousness is non-physical what *does* it show? Robert Howell has suggested that it should not be construed as an argument about ontology of consciousness but about its epistemology. Particularly, the knowledge arguments shows that *objectivism*, the view that everything can be known independently of one's own experiential state, is false. It does not bear directly on the ontology of experiential states themselves. This is why the knowledge argument can be leveled against objectivist physicalists and objectivist dualists alike, that is dualists and physicalists that hold that consciousness can be known independently of one's particular experiential perspective. Some facts can only be known from a particular experiential point of view but this is independent of the question whether the facts known are physical or irreducibly mental. The epistemological point leaves the ontology open.[20]

Anti-objectivism is compatible with the metaphysical anti-realism and the denial of epistemic transcendence. Objectivism put limits on *how* certain things can be

[18] Jackson 1998.
[19] Perry 2001.
[20] Howell 2007. Similar points are made in Crane 2003 and Yanyan 2012. All these are in some degree or another indebted to Horgan 1984.

known, not on what can be known in principle. So while the objectivist anti-realist believes that all that can be known can be known by dispassionate observation and rational inference, the anti-objectivist anti-realist believes that all that can be known can be known by observation, inference, and having certain experiential states.

I fundamentally agree with Howell when he says that the knowledge argument should be regarded primarily as an argument for anti-objectivism. However, I disagree with Howell that this should make us conclude that anti-objectivist *physicalism* is the natural outcome. Rather, I will now argue, nothing settles the question whether consciousness is irreducibly mental or physical.

10.3.2 The Argument for Ontological Indeterminacy

My argument against anti-objectivist physicalism is that, independently of *a priori* entailment relations from the physical to the phenomenal, it is hard to see why our anti-objectivism is still properly *physicalist*. Independently of such a criterion it seems that physicalism has no edge over dualism in explaining phenomenal facts. It makes no substantial difference if we claim that phenomenal facts, which we know cannot be fully derived from the physical, are physical but underiveable from physical facts (what is sometimes called *non-reductive* physicalism) or by claiming that they are irreducibly mental. For the anti-realist this has to to entail that there is no fact of the matter regarding which of these positions is true: Consciousness is ontologically indeterminate.

Howell has argued that physicalism can still be defended on grounds of supervenience of the phenomenal on the physical: It is plausible that every change in phenomenal consciousness necessitates a change in physical processes. So, the thought goes, the anti-objectivist can still be a physicalist because she can claim that consciousness is properly supervenient.[21]

There are however well acknowledged problems with defining physicalism by relying on an unqualified notion of supervenience. For instance, *emergentists* hold that phenomenal properties are genuinely novel properties that spring into existence when systems of a certain complexity form.[22] Evidently, emergentists are dualists that defend the supervenience of the phenomenal on the physical and

[21] Howell 2007.
[22] Broad 1925.

would thus be classified as dualists in the sense of Howell.[23] Thus the objection to anti-objectivist physicalism stands. No clear criterion for why anti-objectivist physicalism properly bears this label has been provided.

One may try solving this problem by explicitly defining physicalism as supervenience together with the qualification that no metaphysically distinct properties from physical properties come into existence at any point. However, to make this view intelligible, we arguably need an account of metaphysical distinctness that does not in turn appeal to *a priori* derivability.[24] The difficulties involved in constructing such an account is precisely the problem!

Another approach may be to hold that the anti-objectivist can still hold that consciousness is physical insofar as it is second-order explainable in terms of physical facts. However it seems that second-order explainability clearly offers too weak a standard to accept any particular fundamental ontology. We accept a fundamental ontology because of what it can explain, not because it makes accurate predictions regarding what it cannot explain. The suggestion to appeal to second-order knowability seems unduly *ad hoc*.

Note that the point here is that it would be an awkward *definition* of physicalism that physicalism holds precisely if physical facts make the mental facts explainable *or* second-order explainable. A metaphysical realist may hold that one can still make a strong case for physicalism based on second-order explainability.[25] On such a view it may be that the physical facts still necessitate the phenomenal facts, though in a epistemically transcendent fashion. But of course the metaphysical anti-realist denies that transcendent necessitation makes sense. So, for her, the entailment from second-order explainability to physicalism would have to be a brute fact. This is an exceedingly implausible interpretation of what it takes to be a physicalist.

A second reason why second-order knowability of the phenomenal is insufficient to argue for physicalism is that the relevant second-order phenomenal knowledge plausibly requires at least *some* first-order phenomenal knowledge. That is, Mary, given she knows that Bayesian representationalism is true, can know that she will make novel experiences upon leaving her room because novel semantic primitives will

[23] Kim 2000; Crane 2010.
[24] Stoljar 2017.
[25] Such a view is defended by adherents of the so called *phenomenal concept strategy*. For an overview and critical discussion, see Chalmers 2006a.

become active in the inferential network of her mind. However, when thinking about these novel experiences, she will do so in analogy to her black and white experiences. Thus, it is not the case that she could infer knowledge of consciousness from some non-conscious external standpoint. It is merely the case that she can infer some facts about novel experiences from her own conscious standpoint. I think this point is not quite conclusive, however it makes a further intuitive case for the view that second-order knowability is an insufficient ground for a stable physicalist view.

Another possibility of spelling out in what sense anti-objectivism is still physicalist would be in terms of *metaphysical supervenience*, where something is metaphysically supervenient on some base if changes in the supervenient set of facts entail changes in the base in all metaphysically possible worlds. Plausibly, metaphysically possible worlds can be spelled out as notional worlds, given knowledge of all metaphysical natures of things. Further, on the metaphysical anti-realists view, metaphysical natures will be knowable.

But why is it plausible that phenomenal properties are metaphysically supervenient on the physical? Arguably, this is again because phenomenal properties are second-order knowable. But if this is so, then the appeal to metaphysical supervenience is only nominally different from the appeal to second-order knowability as a criterion of physicalism and should be rejected on the very same grounds. Second-order knowability, given the physical facts, is much too weak a criterion for deciding for physicalism as a fundamental ontology.

It should by now have become clear that these difficulties are quite systematic. And thus, from the anti-realist vantage point, it seems that as long as there is a fact that cannot be derived from physics alone physicalism offers an incomplete account of reality. And the view that phenomenal facts are somehow physical "deep down", is unintelligible on an anti-realist meta-ontology.

I conclude that physicalism offers no explanatory advantage over dualism in the sense required. For the metaphysical realist, who, at least superficially, seems to be committed to ontological determinacy this would mean that we will never know what the true ontology of consciousness is. In fact, this position has been adopted under the label of *mysterianism*. Mysterians hold that consciousness is

physical but in a way that is in-principle unintelligible to human minds.[26] But for the metaphysical anti-realist this is not an option. If in-principle unknowable explanations are metaphysical fantasies then the only coherent solution is that consciousness is ontologically indeterminate. In a sense, ontological indeterminacy is mysterianism without the mystery. The attempt to understand consciousness in terms of some fundamental ontological postulate springs from the desire to contemplate the world from an absolute perspective. If the argument presented here is somewhat on the right track, then this can never be achieved. The world can ever only be known from within. And when known from within, there is no need to decide for a unified fundamental ontology.

10.4 Summary

Part two motivated a representationalist and functionalist take on consciousness based on Bayesian cognitive science. Now in this part we have engaged with the question whether Bayesian cognitive science can also teach us something about the metaphysics of the conscious mind. If my diagnosis that Bayesian cognitive science is best understood as entailing a sort of metaphysical anti-realism, then the answer is 'yes'. For the various mysteries of consciousness, those that have given rise to the 'hard problem' begin to dissolve once subjected to the metasemantic and meta-ontological constraints of by metaphysical anti-realism.

Model-relative realism says that the only reasonable way of thinking about the world is to do so from within. Bayesian representationalism says that consciousness is an inevitable result of organisms modelling themselves and their environment in an approximately Bayesian fashion. Ontological indeterminacy says that, because the world can only be appreciated from within, certain questions about how subjectivity itself emerges from physical reality are without answers. They are the result of our propensity to reason from God's eye point of view.

[26] McGinn 2011.

Conclusion

Starting from phenomenological observations I argued for representationalism. Consciousness is representational through and through and so understanding consciousness requires understanding the representational relation. To investigate the nature of mental representation we turned to Bayesian cognitive science, which may be the most advanced attempt at understanding how the brain constructs and updates models of its environment in the service of adaptive behavior. Building on these insights, I suggested a potential theory of consciousness that explains qualitative consciousness in terms of perceptual inference in the predictive hierarchy and explains reflexive consciousness in terms of introspective inference.

But Bayesian cognitive science bears on issues of conscious experience in another way, too. On a more speculative note I argued that taking Bayesian cognitive science seriously supports a form of metaphysical anti-realism: Bayesian cognitive science in principle can not explain how we could ever as much as refer to a wholly mind-independent reality. This, I showed in the final chapter, has important implications for the metaphysics and ontology of consciousness because it makes it possible to hold that the ontology of consciousness is indeterminate.

I want to close the thesis by collecting a few general thoughts that came to me while pondering these ideas but that did not find a place within the main discussion. On the scientific side of things, specifically in the discipline of consciousness science, there are two main suggestions that result from my reasoning. First, there is the general representationalist point. For the representationalist, consciousness science is a science of a specific type of mental representations. Specifically, for the self-representationalist, consciousness science should study the nature of reflexive mental representations. It seems to me that, outside of philosophy, the idea that consciousness is really just a specific form of mental representation is only slowly

gaining ground. Also, if I am right to say that there is an ambiguity implicit in our ordinary concept of consciousness, namely one between mere qualitative and reflexive consciousness, then keeping these aspect firmly apart will be relevant to avoiding confusion in empirical research on the matter.

Secondly, if my approach to the hard problem is correct, then consciousness researchers should, to a large extend, not worry about these metaphysical issues. According to Bayesian representationalism, phenomenal qualities result form low-level representational states that serve as semantic primitives with respect to an inferential hierarchy. The above analysis of the knowledge argument suggests that representations of semantic primitives will be associated with irreducible phenomenal knowledge that cannot be conclusively reduced to physical processes. But this lack of reducibility should not be viewed as an invitation to metaphysical speculation by the scientist, but as a place of ontological indeterminacy: No deep mysteries are waiting to be unveiled at this vanishing point of subjectivity.

I also want to offer some thoughts on the relation of Bayesian inferentialism to a particular hobbyhorse of mine, namely the possibility of artificial consciousness. Certainly artificial intelligence will be one of the most vibrant fields of scientific progress in this century. One does not have to be all that prescient to see that the question whether certain systems are actually conscious or whether they are just clever imitations of conscious agency will become more than just an issue of philosophical curiosity quite soon. So what can the approach in this thesis tell us about the plausibility of artificial consciousness.

Maybe the most important point here is to say that the approach to consciousness was fully functionalist. Phenomenal facts are second-order knowable, given the functional facts, and thus supervenient on those facts. Thus nothing but the right abstract causal organization is required for a system to develop consciousness. This is important because it entails that achieving artificial consciousness should at least be possible in principle using computer technology available today, though I do not want to comment on the amount and kind of data-processing that may or may not be necessary to achieve lighting the spark of consciousness.

Further, the thesis that phenomenal consciousness is really an amalgamation of computational features such as the hierarchical representation of low-level semantic

primitives, reflexive self-directed processing and global integration in the service of action guidance should make us weary of asking whether a particular system is or is not conscious *simpliciter*. Conscious in what sense? In the sense of representing its environment in terms of a hierarchical generative model that possesses a lowest layer of semantic primitives? Arguably some artificial systems today fit this bill.

What about reflexivity and global integration? Here we can make a few predictions. If the winning hypothesis account is correct the integration of information into a global neuronal workspace will require that the relevant computational system is geared towards action in a dynamic environment. No mere passive system, like an artificial weather watcher, or maybe a large language model like $GPT3$[27], would necessitate global integration. The processing in such a system would not show the symmetry break between conscious and unconscious processes characteristic of the human psyche. However, as the winning hypothesis account explains integration building on mere self-organization, given that a sufficiently complex computational system actively engages with its environment we should expect the emergence of global integration as a natural consequence.

Finally, what about reflexivity? Above we speculated that reflexive processing is the result of inductive inference to mental causes. Such inferences can be expected to emerge in complex computational systems that are faced with unexpected scenarios. But here too it seems that no deliberate attempt at building reflexivity into a conscious system is necessary. If an adaptive computational system gets sophisticated enough reflexivity should be expected to emerge as a natural consequence.

So it seems we are drawn to the conclusion that consciousness can be expected to naturally emerge in advanced systems that are interacting with a complex environment. Consciousness is a generic feature of complex action-oriented data processing rather than an adaption tied to some peculiar biological makeup. This insight is both elating and humbling. Elating, because it shows that conscious minds aren't mere flukes of chance selection but the almost inevitable outcomes of the evolution of complexity in the universe. Humbling, because of what may lie ahead.

[27]Though this would require further discussion.

List of Figures

Bibliography

Armstrong, David Malet (1968). "The Headless Woman Illusion and the Defence of Materialism". In: *Analysis* 29.2, pp. 48–9.

– (1980). *The Nature of Mind and Other Essays*. University of Queensland Press.

Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Barz, Wolfgang (2019). "The Puzzle of Transparency and How to Solve It". In: *Canadian Journal of Philosophy* 49.7, pp. 916–935.

– (2021). "The Distinct Existences Argument Revisited". In: *Synthese* 199, pp. 8443–8463.

Bayne, Tim (2010). *The Unity of Consciousness*. Oxford University Press.

Bayne, Tim and David Chalmers (2003). "What is the Unity of Consciousness?" In: *The Unity of Consciousness*. Ed. by Axel Cleeremans. Oxford University Press.

Block, Ned (1978). "Troubles with Functionalism". In: *Minnesota Studies in the Philosophy of Science* 9, pp. 261–325.

– (1986). "Advertisement for a semantics for psychology". In: *Midwest Studies in Philosophy* 10, pp. 615–678.

– (1990). "Inverted Earth". In: *Philosophical Perspectives* 4.1990, p. 53.

– (1996). "Mental Paint and Mental Latex". In: *Philosophical Issues* 7.1996, pp. 19–49.

– (2010). "Attention and Mental Paint". In: *Philosophical Issues* 20.1, pp. 23–63.

Block, Ned (2018). "If Perception is Probabilistic, Why Doesn't It Seem Probabilistic?" In: *Philosophical Transactions of the Royal Society B* 373.1755.

Bogacz, Rafal (2017). "A tutorial on the free-energy framework for modelling perception and learning". In: *Journal of Mathematical Psychology* 76, pp. 198–211.

Boghossian, Paul (1989). "The Rule-Following Considerations". In: *Mind* 98.392, pp. 507–49.

Boolos, George and Richard Jeffrey (1980). *Computability and Logic*. 2nd ed. Cambridge University Press.

Brandom, Robert B. (1996). *Making it Explicit*. Harvard University Press.

– (2000). *Articulating Reasons*. Harvard University Press.

Broad, C. D. (1925). *The Mind and its Place in Nature*. Routledge.

Burgess, J. A. (2007). "When is Circularity in Definitions Benign?" In: *Philosophical Quarterly* 58.231, pp. 214–233.

Button, Tim (2013). *The Limits of Realism*. Oxford University Press UK.

Byrne, Alex (2001). "Intentionalism Defended". In: *The Philosophical Review* 110.2, pp. 199–240.

Byrne, Alex and David Hilbert (2011). "Are Colors Secondary Qualities?" In: *Primary and Secondary Qualities: The Historical and Ongoing Debate*. Ed. by Lawrence Nolan. Oxford University Press.

Carhart-Harris, Robin and Karl Friston (Feb. 2010). "The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas". In: *Brain : a journal of neurology* 133, pp. 1265–83.

Carnap, Rudolf (1950). *Logical Foundations of Probability*. Chicago University of Chicago Press.

Chalmers, David (1995). "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3, pp. 200–19.

– (1996). "Does a Rock Implement Every Finite-State Automaton?" In: *Synthese* 108.3, pp. 309–33.

– (1998). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

– (2003a). "The Content and Epistemology of Phenomenal Belief". In: *Consciousness: New Philosophical Perspectives*. Ed. by Quentin Smith and Aleksandar Jokic. Oxford University Press, pp. 220–72.

– (2003b). "The Nature of Narrow Content". In: *Philosophical Issues* 13.1, pp. 46–66.

– (2004). "The Representational Character of Experience". In: *The Future for Philosophy*. Ed. by Brian Leiter. Oxford University Press, pp. 153–181.

– (2006a). “Phenomenal Concepts and the Explanatory Gap”. In: *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Ed. by Torin Alter and Sven Walter. Oxford University Press.
– (2006b). “Two-Dimensional Semantics”. In: *The Oxford Handbook to the Philosophy of Language*. Ed. by E. Lepore and B. Smith. Oxford University Press.
– (2009). “The Two-Dimensional Argument Against Materialism”. In: *Oxford Handbook to the Philosophy of Mind*. Ed. by Brian P. McLaughlin and Sven Walter. Oxford University Press.
– (2010). “Perception and the Fall from Eden”. In: *Perceptual Experience*.
– (2012a). *Constructing the World*. Oxford University Press.
– (2012b). “On implementing a computation”. In: *Machine Intelligence: Perspectives on the Computational Model*. 1990, pp. 109–120.
– (2012c). “The Varieties of Computation: A Reply”. In: *Journal of Cognitive Science* 2012.3, pp. 211–248.
– (2018). “The Meta-Problem of Consciousness”. In: *Journal of Consciousness Studies* 25.9-10, pp. 6–61.
– (2021). “Inferentialism, Australian Style”. In: *Proceedings and Addresses of the American Philosophical Association* 92.

Churchland, Paul M. (1989). “Knowing Qualia: A Reply to Jackson”. In: *A Neurocomputational Perspective*. Ed. by Yujin Nagasawa, Peter Ludlow, and Daniel Stoljar. MIT Press, pp. 163–178.

Clark, Andy (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
– (2017). “Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil”. In: *Noûs* 51.4, pp. 727–753.

Clark, Andy, Karl Friston, and Sam Wilkinson (2019). “Bayesing Qualia”. In: *Journal of Consciousness Studies* 26.9, pp. 19–33.

Colyvan, Mark (2003). “The Philosophical Significance of Cox’s Theorem”. In: *International Journal of Approximate Reasoning* 37.

Conee, Earl (1987). “Reason, Truth and History”. In: *Noûs* 21.1, pp. 81–95.

Cowie, Fiona (2008). *Innateness and Language*. `https://plato.stanford.edu/entries/innateness-language/`. Accessed: 2022-09-20.

Cox, Richard T. (1946). “Probability, Frequency, and Reasonable Expectation”. In: *American Journal of Physics* 14.2, pp. 1–13.

Crane, Tim (2003). “Subjective Facts”. In: *Real Metaphysics*. Ed. by Hallvard Lillehammer and Gonzalo Rodriguez Pereyra. London: Routledge, pp. 68–83.

– (2010). “Cosmic Hermeneutics Vs. Emergence: The Challenge of the Explanatory Gap”. In: *Emergence in Mind*. Ed. by Cynthia Macdonald and Graham Macdonald. Oxford: Oxford University Press, pp. 22–34.

Dalen, Dirk van (2008). *Logic and Structure*. 2nd ed. Springer.

Dehaene, Stanislas (2011). “Conscious and Nonconscious Processes: Distinct Forms of Evidence Accumulation?” In: *Biological Physics: Poincaré Seminar 2009*. Ed. by Vincent Rivasseau. Basel: Springer Basel, pp. 141–168.

– (2014). *Consciousness and the Brain. Deciphering How the Brain Codes Our Thoughts*. Viking Press.

Dehaene, Stanislas and Lionel Naccache (2001). “Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework”. In: *Cognition* 79.1. The Cognitive Neuroscience of Consciousness, pp. 1–37.

Dennett, Daniel C. (1988). “Quining Qualia”. In: *Consciousness in Modern Science*. Ed. by Anthony J. Marcel and E. Bisiach. Oxford University Press.

– (1991). *Consciousness Explained*. Penguin Books.

– (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Norton.

Dennett, Daniel C. and Marcel Kinsbourne (1992). “Time and the observer: The where and when of consciousness in the brain”. In: *Behavioral and Brain Sciences* 15.2, pp. 183–201.

Derrien, Diane et al. (Feb. 2022). “The nature of blindsight: implications for current theories of consciousness”. In: *Neuroscience of Consciousness* 2022.1. niab043.

Devitt, Michael (1983). “Realism and the Renegade Putnam: A Critical Study of Meaning and the Moral Sciences”. In: *Noûs* 17.2, pp. 291–301.

Dołega, Krzysztof and Joe E. Dewhurst (2021). “Fame in the Predictive Brain: A Deflationary Approach to Explaining Consciousness in the Prediction Error Minimization Framework”. In: *Synthese* 198.8, pp. 7781–7806.

Douven, Igor (1999). “Putnam?s Model-Theoretic Argument Reconstructed”. In: *Journal of Philosophy* 96.9, pp. 479–490.

Douven, Igor, Leon Horsten, and Jan-Willem Romeijn (2010). "Probabilist Antirealism". In: *Pacific Philosophical Quarterly* 91.1, pp. 38–63.

Dretske, Fred (1986). *Knowledge and the Flow of Information*. Vol. 60. 1, pp. 116–121.

– (1995). *Naturalizing the Mind*. MIT Press.

Dummett, Michael (1975). "The Philosophical Basis of Intuitionistic Logic". In: *Truth and Other Enigmas*. Ed. by Michael Dummett. Cambridge: Harvard UP, pp. 215–247.

– (1991). *The Logical Basis of Metaphysics*. Harvard University Press.

– (2001). "Victor's Error". In: *Analysis* 61.1, pp. 1–2.

Eckardt, Barbara Von (2012). "The Representational Theory of Mind". In: *The Cambridge Handbook of Cognitive Science*. Ed. by Keith Frankish and William Ramsey. Cambridge University Press.

Edgington, Dorothy (1985). "The Paradox of Knowability". In: *Mind* 94.376, pp. 557–568.

Egan, Andy (2006). "Appearance Properties?" In: *Noûs* 40.3, pp. 495–521.

Es, Thomas van (2021). "Living models or life modelled? On the use of models in the free energy principle". In: *Adaptive Behavior* 29.3, pp. 315–329.

Es, Thomas van and Erik Myin (2020). "Predictive Processing and Representation: How Less Can Be More". In: *The Philosophy and Science of Predictive Processing*. Ed. by Dina Mendonça, Manuel Curado, and Steven S. Gouveia, pp. 7–24.

Feldman, Harriet and Karl Friston (Dec. 2010). "Attention, Uncertainty, and Free-Energy". In: *Frontiers in human neuroscience* 4.

Fine, Terence L. (1973). *Theories of Probability*. Academic Press.

Fitch, Frederic (1963). "A Logical Analysis of Some Value Concepts". In: *Journal of Symbolic Logic* 28.2, pp. 135–142.

Frankish, Keith (2016). "Illusionism as a Theory of Consciousness". In: *Journal of Consciousness Studies* 23.11-12, pp. 11–39.

Franklin, J. (2001). "Resurrecting Logical Probability". In: *Erkenntnis* 55.2, pp. 277–305.

Frege, Gottlob (1892). "Über Sinn Und Bedeutung". In: *Zeitschrift für Philosophie Und Philosophische Kritik* 100.1, pp. 25–50.

Freud, Sigmund (1966). *Die Traumdeutung*. Fischer.

Friston, Karl (2005). “A theory of cortical responses”. In: *Philosophical Transactions of the Royal Society B* 360.1456, pp. 815–836.

– (2010). “The free-energy principle: A unified brain theory?” In: *Nature Reviews Neuroscience* 11.2, pp. 127–138.

Friston, Karl and Ping Ao (2012). “Free energy, value, and attractors”. In: *Computational and Mathematical Methods in Medicine*.

Friston, Karl, Wanja Wiese, and J. Allan Hobson (2020). “Sentience and the origins of consciousness: From cartesian duality to Markovian monism”. In: *Entropy* 22.5, pp. 1–31.

Frith, Chris (2007). *Making up the Mind: How the Brain Creates our Mental World*. Blackwell Publishing.

Gennaro, Rocco J. (2006). “Between Pure Self-Representationalism and the Extrinsic HOT Theory of Consciousness”. In: *Self-Representational Approaches to Consciousness*. Ed. by Uriah Kriegel and Kenneth Williford. MIT Press.

Gentaz, Edouard et al. (June 2004). “The Visual and the Haptic Müller-Lyer Illusions: Correlation Study”. In: *Current psychology letters*.

Gentzen, Gerhard (1935). “Untersuchungen Über Das Logische Schließen. I.” In: *Mathematische Zeitschrift* 35, pp. 176–210.

Gładziejewski, Paweł (2016). “Predictive coding and representationalism”. In: *Synthese* 193.2, pp. 559–582.

Gładziejewski, Paweł and Miłkowski, Marcin (2017). “Structural representations: causally relevant and different from detectors”. In: *Biology and Philosophy* 32.3, pp. 337–355.

Godfrey-Smith, Peter (2004). “Mental Representation, Naturalism, and Teleosemantics”. In: *Teleosemantics: New Philosophical Essays*. Ed. by David Papineau and Graham MacDonald. Oxford University Press, pp. 42–68.

– (2009). “Triviality Arguments Against Functionalism”. In: *Philosophical Studies* 145.2, pp. 273–295.

Hájek, Alan (2019). *Interpretations of Probability*. `https://plato.stanford.edu/entries/probability-interpret/`. Accessed: 2021-12-06.

Harman, Gilbert (1987). "(Nonsolipsistic) Conceptual Role Semantics". In: *New Directions in Semantics*. Ed. by Ernest LePore. London: Academic Press, pp. 55–81.

– (1990). "The Intrinsic Quality of Experience". In: *Philosophical Perspectives* 4, p. 31.

Haukioja, Jussi (2017). "Postscript: Recent Work on Putnam's Model-Theoretic Argument". In: *A Companion to Philosophy of Language*. Ed. by Bob Hale. Blackwell, pp. 730–733.

Hawkins, Jeffrey (2004). *On Intelligence*. Owl Books.

Heidegger, Martin (1927). *Sein Und Zeit*. M. Niemeyer.

Helmholtz, Hermann v. (1921). "Die Tatsachen in der Wahrnehmung". In: *Schriften zur Erkenntnistheorie*. Ed. by Paul Hertz and Moritz Schlick. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 109–175.

Hilbert, David R. and Mark Eli Kalderon (2000). "Color and the Inverted Spectrum". In: *Vancouver Studies in Cognitive Science*. Ed. by Steven Davis. New York: Oxford University Press, pp. 187–214.

Hipólito, Inês, Maxwell Ramstead, and Karl Friston (2020). "Is the Free-Energy Principle a Formal Theory of Semantics? From Variational Density Dynamics to Neural and Phenotypic Representations". In: *Entropy* 22.8, pp. 1–30.

Hoffman, Donald (2020). *The Case Against Reality: How Evolution hid the Truth from our Eyes*. Penguin Books.

Hohwy, Jakob (Apr. 2012). "Attention and Conscious Perception in the Hypothesis Testing Brain". In: *Frontiers in psychology* 3, p. 96.

– (2014). *The Predictive Mind*. Oxford University Press.

– (2016). "The Self-Evidencing Brain". In: *Noûs* 50.2, pp. 259–285.

– (2020). "Self-Supervision, Normativity and the Free Energy Principle". In: *Synthese* 199.1-2, pp. 29–53.

Horgan, Terence (1984). "Jackson on Physical Information and Qualia". In: *Philosophical Quarterly* 34.April, pp. 147–52.

Horgan, Terence and John Tienson (2002). "The Intentionality of Phenomenology and the Phenomenology of Intentionality". In: *Philosophy of Mind: Classical and Contemporary Readings*. Ed. by David Chalmers. Oup Usa, pp. 520–533.

Horgan, Terence, John Tienson, and Graham George (2006). "Internal-World Skepticism and the Self-Presentational Nature of Phenomenal Consciousness". In: *Self-representational Approaches to Consciousness*. Ed. by Kriegel Uriah and Kenneth Williford. Bradford.

Horgan, Terry and Uriah Kriegel (2007). "Phenomenal Epistemology: What is Consciousness That We May Know It so Well?" In: *Philosophical Issues* 17.1, pp. 123–144.

Howell, Robert J. (2007). "The Knowledge Argument and Objectivity". In: *Philosophical Studies* 135.2, pp. 145–177.

Huemer, Michael (2001). *Skepticism and the Veil of Perception*. Lanham: Rowman & Littlefield.

Humberstone, I. L. (1997). "Two Types of Circularity". In: *Philosophy and Phenomenological Research* 57.2, pp. 249–280.

Jackson, Frank (1986). "What Mary Didn't Know". In: *Journal of Philosophy* 83.5, pp. 291–295.

– (1998). *Mind, Method and Conditionals: Selected Papers*. Routledge.

– (2003). "Mind and Illusion". In: *Royal Institute of Philosophy Supplement* 53.June, pp. 251–271.

James, William (1950). *Principles of Psychology: Volume One*. Dover.

Jaynes, Edwin T. (1957). "Information Theory and Statistical Mechanics. II". In: *Physical Review* 108.2, p. 171.

– (1988). "How does the Brain Do Plausible Reasoning?" In: *Maximum-Entropy and Bayesian Methods in Science and Engineering. Fundamental Theories of Physics*. Ed. by Smith C.R. Erickson G.J. Springer.

– (2003). *Probability Theory. The Logic of Science*. Cambridge University Press.

Jaynes, Julian (1976). "The Origin of Consciousness in the Breakdown of the Bicameral Mind". In: *Philosophy and Rhetoric* 14.2, pp. 127–129.

Jeffrey, Richard C. (1990). *The Logic of Decision*. 2nd ed. Univerysity of Chicago Press.

Joyce, James M. (2004). "Bayesianism". In: *The Oxford Handbook of Rationality*. Ed. by Piers Rawling and Alfred R. Mele. Oxford: Oxford University Press, pp. 132–155.

Jung, Carl Gustav (1971). *Zwei Schriften über analytische Psychologie*. Walter.

Kaufmann, Rafael, Pranav Gupta, and Jacob Taylor (2021). "An active inference model of collective intelligence". In: *Entropy* 7.23.

Keynes, John Maynard (1921). *A Treatise on Probability*. London, England: Dover Publications.

Kiefer, Alex and Jakob Hohwy (2018). "Content and misrepresentation in hierarchical generative models". In: *Synthese* 195.6, pp. 2387–2415.

– (2019). "Representation in the Prediction Error Minimization Framework". In: *The Routledge Companion to Philosophy of Psychology: 2nd Edition*. Ed. by Sarah K. Robins, John Symons, and Paco Calvo, pp. 384–409.

Kim, Jaegwon (2000). *Mind in a Physical World. An Essay on the Mind-Body Problem and Mental Causation*. MIT-Press.

Kind, Amy (2003). "What's so Transparent About Transparency?" In: *Philosophical Studies* 115.3, pp. 225–244.

Klein, Colin (2008). "Dispositional Implementation Solves the Superfluous Structure Problem". In: *Synthese* 165.1, pp. 141–153.

Kriegel, Uriah (2002). "Phenomenal Content". In: *Erkenntnis* 57.2, pp. 175–198.

– (2008). "Real narrow content". In: *Mind and Language* 23.3, pp. 304–328.

– (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford University Press.

– (2011). *The Sources of Intentionality*. Oxford University Press, pp. 1–288.

– (2017). "Reductive Representationalism and Emotional Phenomenology". In: *Midwest Studies in Philosophy* 41.1, pp. 41–59.

Kripke, Saul (1980). *Naming and Necessity*. Harvard University Press.

Leeds, Stephen (1993). "Qualia, Awareness, Sellars". In: *Noûs* 27.3, pp. 303–330.

Lepore, Ernest and Jerry Fodor (1993). "Précis of Holism: A Shopper's Guide". In: *Philosophy and Phenomenological Research* 53.3, pp. 637–640.

Lewis, David (1970). "How to Define Theoretical Terms". In: *Journal of Symbolic Logic* 36.2, pp. 427–446.

– (1983). "Postscript to "Mad Pain and Martian Pain"". In: *Philosophical Papers* I, pp. 130–132.

– (1984). “Putnam’s Paradox”. In: *Australasian Journal of Philosophy* 62.3, pp. 221–236.

Lycan, William G. (1996). *Consciousness and Experience*. MIT Press.

MacKay, David J. C. (2002). *Information Theory, Inference Learning Algorithms*. USA: Cambridge University Press.

Marvan, Tomas and Marek Havlík (2021). “Is Predictive Processing a Theory of Perceptual Consciousness?” In: *New Ideas in Psychology* 61.21.

Mashour, George et al. (Mar. 2020). “Conscious Processing and the Global Neuronal Workspace Hypothesis”. In: *Neuron* 105, pp. 776–798.

Maudlin, Tim (1989). “Computation and Consciousness”. In: *Journal of Philosophy* 86.8, pp. 407–432.

McClelland, Tom (forthcoming). “Self-Representationalism”. In: *The Oxford Handbook of the Philosophy of Consciousness*. Ed. by Uriah Kriegel. Oxford: Oxford University Press.

McDowell, John (1985). “Values and Secondary Qualities”. In: *Morality and Objectivity*. Ed. by Ted Honderich. London: Routledge, pp. 110–129.

McGinn, Colin (1984). “The Subjective View”. In: *Philosophy* 59.228, pp. 272–275.

– (1996). “Another Look at the Colors”. In: *Journal of Philosophy* 93.11, pp. 537–553.

– (1997). “Fred Dretske’s Naturalizing the Mind (MIT Press, 1995) Missing the Mind: Consciousness in the Swamps”. In: *Noûs* 31.4, pp. 528–537.

– (2011). *The Mysterious Flame*. Basic Books.

Mendelovici, Angela (2018). *The Phenomenal Basis of Intentionality*. Oxford University Press.

Mendelovici, Angela and David Bourget (2014). “Naturalizing Intentionality: Tracking Theories Versus Phenomenal Intentionality Theories”. In: *Philosophy Compass* 9.5, pp. 325–337.

Metzinger, Thomas (2009). *The Ego Tunnel: The Science of Mind and the Myth of the Self*. Basic Books.

Millidge, Beren, Alexander Tschantz, and Christopher Buckley (Jan. 2021). “Whence the Expected Free Energy?” In: *Neural Computation* 33, pp. 1–36.

Millikan, Ruth Garrett (1989). “Biosemantics”. In: *The Journal of Philosophy* 86.6, p. 281.

Moore, G. E. (1903). “The Refutation of Idealism”. In: *Mind* 12.48, pp. 433–453.

Mossio, Matteo, Cristian Saborido, and Alvaro Moreno (2009). “An Organizational Account of Biological Functions”. In: *British Journal for the Philosophy of Science* 60.4, pp. 813–841.

Nagel, Thomas (1971). “Brain Bisection and the Unity of Consciousness”. In: *Synthese* 22.3/4, pp. 396–413.

– (1974). “What Is It Like to Be a Bat?” In: *The Philosophical Review* 83.4, pp. 435–450.

– (2012). *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. Oxford Up.

Neander, Karen (2017). *A Mark of the Mental: A Defence of Informational Teleosemantics*. Cambridge, USA: MIT Press.

Newman, M. H. A. (1928). “Mr. Russell’s Causal Theory of Perception”. In: *Mind* 37.146, pp. 26–43.

Nisbett, Richard E. and Timothy D. Wilson (1977). “Telling More Than We Can Know: Verbal Reports on Mental Processes”. In: *Psychological Review* 84.3, pp. 231–59.

O’Brien, Gerard and Jon Opie (2004). “Notes Toward a Structuralist Theory of Mental Representation”. In: *Representation in Mind*, pp. 1–20.

Paris, J. B. (1994). *The Uncertain Reasoner’s Companion: A Mathematical Perspective*. Cambridge University Press.

Parr, Thomas, Giovanni Pezzulo, and Karl Friston (2022). *Active Inference. The Free Energy Principle in Mind, Brain and Behavior*. MIT-Press.

Pautz, Adam (2006). “Sensory awareness is not a wide physical relation: An empirical argument against externalist intentionalism”. In: *Nous* 40.2, pp. 205–240.

– (2010). “Do Theories of Consciousness Rest on a Mistake?” In: *Philosophical Issues* 20.1, pp. 333–367.

Pautz, Adam (2021). “Consciousness Meets Lewisian Interpretation Theory: A Multistage Account of Intentionality”. In: *Oxford Studies in Philosophy of Mind*. Ed. by Uriah Kriegel.

Pearl, Judea (2001). “Causal Inference in Statistics : A Gentle Introduction”. In: *Science* 33.August, pp. 1–20.

Peirce, Charles Sanders (1868). “Questions Concerning Certain Faculties Claimed for Man”. In: *Journal of Speculative Philosophy* 2, pp. 103–114.

Perry, John (2001). *Knowledge, Possibility, and Consciousness*. MIT Press.

Piccinini, Gualtiero (2007). “Computing Mechanisms”. In: *Philosophy of Science* 74.4, pp. 501–526.

Popper, K. R. and J. Eccles (1977). *The Self and its Brain*. Springer International.

Putnam, Hilary (1975). “The Meaning of ’Meaning’”. In: *Minnesota Studies in the Philosophy of Science* 7, pp. 131–193.

– (1977). “Realism and Reason”. In: *Proceedings and Addresses of the American Philosophical Association* 50.6, pp. 483–498.

– (1981). *Reason, Truth and History*. Cambridge University Press.

– (1982). “Why There Isn’t a Ready-Made World”. In: *Synthese* 51.2, pp. 205–228.

– (1987). *Representation and Reality*. MIT Press.

– (1990). *Realism with a Human Face*. Harvard University Press.

– (1993). “Realism Without Absolutes”. In: *International Journal of Philosophical Studies* 1.2, pp. 179–192.

– (1994). “Sense, Nonsense, and the Senses: An Inquiry Into the Powers of the Human Mind”. In: *Journal of Philosophy* 91.9, pp. 445–517.

Quine, Willard V. O. (1951). “Two Dogmas of Empiricism”. In: *Philosophical Review* 60.1, pp. 20–43.

Ramsey, Frank P. (1926). “Truth and Probability”. In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R. B. Braithwaite. McMaster University Archive for the History of Economic Thought. Chap. 7, pp. 156–198.

Ramstead, Maxwell JD, Michael D Kirchhoff, and Karl J Friston (2020). “A tale of two densities: active inference is enactive inference”. In: *Adaptive Behavior* 28.4, pp. 225–239.

Rao, Rajesh and Dana Ballard (Feb. 1999). “Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects”. In: *Nature neuroscience* 2, pp. 79–87.

Rey, Georges (1991). “Sensations in a Language of Thought”. In: *Philosophical Issues* 1, pp. 73–112.

Rosenthal, David M. (1986). “Two Concepts of Consciousness”. In: *Philosophical Studies* 49.May, pp. 329–59.

Rudrauf, David et al. (2017). “A mathematical model of embodied consciousness”. In: *Journal of Theoretical Biology* 428, pp. 106–131.

Sartre, Jean Paul (1982). “Die Imagination”. In: *Die Transzendenz des Egos. Philosophische Essays 1931-1939*. Rowohlt, pp. 97–254.

Sayre, Kenneth (1976). *Cybernetics and the Philosophy of Mind*. Routledge and Kegan Paul.

Schrödinger, Erwin (1958). *Mind and Matter*. Cambridge University Press.

Schwarz, Wolfgang (2018). “Imaginary Foundations”. In: *Ergo, an Open Access Journal of Philosophy* 5.20200916, pp. 1–26.

Schwengerer, Lukas (2019). “Self-Knowledge in a Predictive Processing Framework”. In: *Review of Philosophy and Psychology* 10.3, pp. 563–585.

Schwitzgebel, Eric (2006). “The Unreliability of Naive Introspection”. In: *Philosophical Review* 117.2, pp. 245–273.

– (2015). “If Materialism is True, the United States is Probably Conscious”. In: *Philosophical Studies* 172.7, pp. 1697–1721.

– (2020). “Is There Something It?s Like to Be a Garden Snail”. In: *Philosophical Topics* 48.1, pp. 39–63.

Seager, William and David Bourget (Mar. 2017). “Representationalism about Consciousness”. In: pp. 272–287.

Seager, William E. (2006). “The ’Intrinsic Nature’ Argument for Panpsychism”. In: *Journal of Consciousness Studies* 13.10-11, pp. 129–145.

Searle, John R. (1980). “Minds, brains, and programs”. In: *Behavioral and Brain Sciences* 3.3, pp. 417–424.

– (1983). *Intentionality*. Oxford University Press.

– (1984). *Minds, Brains and Science*. Harvard University Press.

– (1994). *The Rediscovery of the Mind*. MIT Press.

– (2000). “Consciousness”. In: *Annual Review of Neuroscience* 23.1, pp. 557–578.

Seidenfeld, Teddy (1986). “Entropy and Uncertainty”. In: *Philosophy of Science* 53.4, pp. 467–491.

Sellars, Wilfrid (1953). “Inference and Meaning”. In: *Mind* 62.247, pp. 313–338.

– (1956). “Empiricism and the Philosophy of Mind”. In: *Minnesota Studies in the Philosophy of Science* 1, pp. 253–329.

Seth, Anil and Karl Friston (2016). “Active interoceptive inference and the emotional brain”. In: *Philosophical Transactions of the Royal Society B*.

Shannon, C. E. (1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3, pp. 379–423.

Shea, Nicholas (2014). “Exploitable Isomorphism and Structural Representation”. In: *Proceedings of the Aristotelian Society* 114, pp. 123–144.

– (2018). *Representation in Cognitive Science*. Oxford University Press.

Shoemaker, Sydney (1994a). “Phenomenal Character”. In: *Noûs* 28.1, pp. 21–38.

– (1994b). “Self-Knowledge and ”Inner Sense”: Lecture I: The Object Perception Model”. In: *Philosophy and Phenomenological Research* 54.2, pp. 249–269.

– (2003). “Content, Character, and Color”. In: *Philosophical Issues* 13.1, pp. 253–78.

Smith, Michael A. (1986). “Peacocke on Red and Red”. In: *Synthese* 68.September, pp. 559–576.

Smith, Ryan, Maxwell J. D. Ramstead, and Alex Kiefer (2022). “Active Inference Models Do Not Contradict Folk Psychology”. In: *Synthese* 200.2, pp. 1–37.

Smithies, Declan (2019). *The Epistemic Role of Consciousness*. Oxford University Press.

Solms, Mark (2021). *The Hidden Spring. A Journey to the Source of Consciousness*. Profile Books.

Speaks, Jeff (2009). “Transparency, Intentionalism, and the Nature of Perceptual Content”. In: *Philosophy and Phenomenological Research* 79.3, pp. 539–573.

– (n.d.). “A Quick Argument Against Phenomenism, Fregeanism, Appearance Property-Ism and (Maybe) Functionalism About Perceptual Content”.

Stalnaker, Robert C. (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press UK.

Steinberger, Florian and Julien Murzi (2017). “Inferentialism”. In: *Blackwell Companion to Philosophy of Language*. Wiley Blackwell, pp. 197–224.

Steinhoff, Gordon (1986). “Internal Realism, Truth and Understanding”. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1986, pp. 352–363.

Stoljar, Daniel (2004). “The Argument from Diaphanousness”. In: *Canadian Journal of Philosophy Supplementary Volume* 30.October 2014, pp. 341–390.

– (2017). *Physicalism*. `https://plato.stanford.edu/archives/win2017/entries/physicalism/`. Accessed: 2022-05-11.

Strawson, Galen (1996). *Mental Reality*. 2nd ed. MIT-Press.

Swanson, Link R. (2016). “The Predictive Processing Paradigm Has Roots in Kant”. In: *Frontiers in Systems Neuroscience* 10, p. 79.

Taylor, Barry (1991). “’Just More Theory’: A Manoeuvre in Putnam’s Model-Theoretic Argument for Antirealism”. In: *Australasian Journal of Philosophy* 69.2, pp. 152–166.

Tennant, Neil (1997). *The Taming of the True*. Oxford University Press.

– (2001). “Is Every Truth Knowable? Reply to Williamson”. In: *Ratio* 14.3, pp. 263–280.

Thau, Michael (2002). *Consciousness and Cognition: Unified Account*. Oxford University Press.

Thomasson, Amie L. (2003). “Introspection and Phenomenological Method”. In: *Phenomenology and the Cognitive Sciences* 2.3, pp. 239–254.

Thompson, Brad J. (2008). “Representationalism and the Argument From Hallucination”. In: *Pacific Philosophical Quarterly* 89.3, pp. 384–412.

Tononi, Giulio et al. (2016). “Integrated Information Theory: From Consciousness to Its Physical Substrate”. In: *Nature Reviews Neuroscience* 17.7, pp. 450–461.

Tye, Michael (1995). *Ten problems of consciousness : a representational theory of the phenomenal mind*. MIT Press.

– (2003). *Consciousness and Persons. Unity and Identity*. MIT Press.

– (2008). *Consciousness Revisited: Materialism Without Phenomenal Concepts*. MIT Press.

Tye, Michael (2014). “Transparency, qualia realism and representationalism”. In: *Educational Technology Research and Development,* 170.1, pp. 39–57.

Van Horn, Kevin S (2003). “Constructing a logic of plausible inference: a guide to Cox’s theorem”. In: *International Journal of Approximate Reasoning* 34.1, pp. 3–24.

Velmans, Max (1991). “Is Human Information Processing Conscious?” In: *Behavioral and Brain Sciences* 14.4, pp. 651–69.

Vlerick, Michael (2014). “Biologising? Putnam: Saving the Realism in Internal Realism”. In: *South African Journal of Philosophy* 33.3, pp. 271–283.

Walsh, Kevin, David McGovern, et al. (Mar. 2020). “Evaluating the neurophysiological evidence for predictive processing as a model of perception”. In: *Annals of the New York Academy of Sciences* 1464.

Walsh, Sean and Tim Button (2018). *Philosophy and Model Theory*. Oxford, UK: Oxford University Press.

Weiskrantz, Lawrence (2007). “The Case of Blindsight”. In: *The Blackwell Companion to Consciousness*. Ed. by Max Velmans and Susan Schneider. Blackwell.

Whyte, Christopher and Ryan Smith (2020). “The predictive global neuronal workspace: A formal active inference model of visual consciousness”. In: *Progress in Neurobiology* 199.

Whyte, Christopher J. (2019). “Integrating the Global Neuronal Workspace Into the Framework of Predictive Processing: Towards a Working Hypothesis”. In: *Consciousness and Cognition* 73, p. 102763.

Wiese, Wanja (2018). *Experiences Wholeness. Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. MIT Press.

Wiese, Wanja and Thomas Metzinger (2017). “Vanilla PP for Philosophers: A Primer on Predictive Processing”. In: *Philosophy and Predictive Processing*, pp. 1–18.

Williamson, Jon (2010). *In Defence of Objective Bayesianism*. Oxford University Press.

Williamson, Timothy (1987). “On the Paradox of Knowability”. In: *Mind* 96.382, pp. 256–261.

– (2000). “Tennant on Knowable Truth”. In: *Ratio* 13.2, pp. 99–114.

Williford, Kenneth et al. (2018). “The Projective Consciousness Model and Phenomenal Selfhood”. In: *Frontiers in Psychology* 9.

Wittgenstein, Ludwig (2016). *Werkausgabe Band 1*. 22nd ed. Suhrkamp Verlag.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yanyan, Zhao (2012). “The Knowledge Argument Against Physicalism: Its Proponents and Its Opponents”. In: *Frontiers of Philosophy in China* 7.2, pp. 304–316.

Zahavi, Dan (2018). “Brain, Mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism”. In: *Husserl Studies* 34.1, pp. 47–61.

Zusammenfassung

In dieser Dissertation wird eine naturalistische Theorie phänomenalen Bewusstseins entworfen. Die Arbeit ist in drei wesentliche Schritte gegliedert. Zuerst wird eine Form des Repräsentationalismus verteidigt, nach dem bewusste Zustände mentale Repräsentationen sind. Im Anschluss wird der repräsentationale Gehalt bewusster Zustände eingehend phänomenologisch analysiert.

Der Repräsentationalismus impliziert, dass eine naturalistische Theorie des Bewusstseins eine naturalistische Theorie mentaler Repräsentationen voraussetzt. Im zweiten Teil der Arbeit, nach einer Kritik klassischer referentialistischer Theorien der Repräsentation, wird ein Inferentialismus entwickelt, welcher besagt, dass der repräsentationale Gehalt mentaler Zustände sich aus ihrer kausalen Rolle in inferentiellen Prozessen ergibt. Des Weiteren wird argumentiert, dass sich diese inferentiellen Prozesse, im Anschluss an zeitgenössische Strömungen in den Kognitionswissenschaften, als approximierte bayesianische Inferenzen verstehen lassen. Insbesondere wird argumentiert, dass phänomenale Eigenschaften sich aus einer genau bestimmbaren Rolle im inferentiellen Geflecht der Kognition erklären lassen. Eine weitere zentrale Aussage des zweiten Teils der Dissertation ist, dass diese philosophische Theorie des Bewusstseins nicht nur die phänomenologischen Beobachtungen aus Teil eins einfangen kann, sondern auch gut mit Erkenntnissen aus der wissenschaftlichen Erforschung des Bewusstseins kompatibel ist.

Der dritte Teil der Arbeit widmet sich metaphysischen Fragen, die sich aus der vermeintlichen Naturalisierung des Bewusstseins ergeben. Es wird hier die These aufgestellt, dass das bayesianische Paradigma in den Kognitionswissenschaften eine gewisse Abschwächung des metaphysischen Realismus, der Position, dass unsere Wahrnehmungen und Gedanken sich auf eine gänzlich geistunabhängige Wirklichkeit beziehen, erfordert. Vielmehr ist der ganze Apparat approximierter bayesianischer

Zusammenfassung

Inferenz nur relativ zu einem vorausgesetzten Modell der Wirklichkeit sinnvoll. Hieraus folgt, wenn wir das bayesianische Paradigma beim Wort nehmen, dass auch die Wirklichkeit, auf welche wir uns beziehen, relativ zu einem Modell gedacht werden muss. Das letzte Kapitel der Arbeit argumentiert, dass diese Abschwächung des metaphysischen Realismus neue Möglichkeiten eröffnet, die Relation von phänomenalem Bewusstsein zu physischen Prozessen zu denken.