UNIVERSITY OF CALIFORNIA
RIVERSIDE

Reduce Them All: A General Theory of Psychophysical Reduction

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Jeremy Michael Pober

September 2022

Dissertation Committee:
Dr. Eric Schwitzgebel, Chairperson
Dr. Michael Nelson
Dr. John Martin Fischer
Dr. Luca Ferrero

The Dissertation of Jeremy Michael Pober is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This work has benefitted from conversations with a great number of people across various topics over the course of several years. I doubt I can recall every one; any omissions here reflect my imperfect memory and not the value of the contributions of those I fail to mention.

For discussions in philosophy of biology, on natural kinds and mechanistic explanation, respectively, Matt Barker and Andrew Bollhagen. For discussions on the metaphysical side of reduction, Kevin Morris. For discussions on functionalism and reduction in philosophy of mind, Andrew MacDonald, Eric Mandelbaum, Chris Masciari, Carlos Montemayor, Jake Quilty-Dunn, J. Brendan Ritchie, Mason Westfall, and the audiences at the 2019 Berkeley-Stanford-Davis Graduate Conference and the 4th 'Cogtweeto' workshop. For introducing me to these issues (too many) years ago, Adam Kovach, Georges Rey, and Ken Richman. For securing me exceptional support from UCR and the Department of Philosophy, Coleen Macnamara (Director of Graduate Studies) and Andrews Reath (Chair). For their time, thoughts, and feedback throughout my time at UCR, Agnieszka Jaworska and my committee members Luca Ferrero and John Martin Fischer. Special thanks to Michael Nelson for dozens, if not hundreds, of hours spent discussing the whole gamut of topics within the metaphysics of mental states, and for pressing me on my views throughout. And finally, to Eric Schwitzgebel, my adviser and chair, for help and support of a degree that I simply cannot convey in the lines of an acknowledgement section.

Dedication

For Neil—you set me on this path; I wish you were still here to walk it with me.

ABSTRACT OF THE DISSERTATION

Reduce Them All: A General Theory of Psychophysical Reduction

by

Jeremy Michael Pober

Doctor of Philosophy, Graduate Program in Philosophy
University of California, Riverside, September 2022
Dr. Eric Schwitzgebel, Chairperson

I propose and defend a novel version of reductive functionalism about mental phenomena. This theory makes three key contributions. First, I argue that reduction should be understood as a question of 'nomic equivalence' between tokens of the reducing and base type, as well as the inheritance of causal powers by the former from the latter. This definition sets aside issues about, e.g., the priority of physics over other special sciences. Second, I suggest a solution to what I call the generality problem. This problem is that other such theories offer only 'local' or species-specific reductions, such that human pain reduces to physical type H, dog pain to physical type D, and so on, but fail to explain what all pain states have in common. I argue that the Homeostatic Property Cluster theory of natural kinds offers a framework in which both more finely grained, type-reducible mental categories as well as broader, species-general mental categories count as natural kinds. Finally, I add a tool to the reductionist's kit for responding to the multiple realizability objection. I propose a novel functionalism in which mental states

are the core, rather than the total realizers of functional roles. This allows the reductionist to handle a distinct class of multiple realizers: those where the core component is embedded differently in distinct realizing mechanisms. For instance, digital video displays are multiply realizable in (at least) LCD and CRT form, but such displays share a core component that is ‘doing the work’ of displaying videos by illuminating various pixels. By identifying the functional (and in this case artifactual) kind with the core component of pixel illuminators, we can reduce it to the sufficiently similar physical bases of those illuminators.

Table of Contents

Chapter One / Introduction

I propose and articulate a metaphysics of mental states that is, broadly speaking, within the family of reductive physicalism. Pinning down exactly what ‘reductive physicalism’ is and/or implies is a delicate affair, and has not, in my opinion, been given enough attention. I therefore dedicate the first section (chapters 2-4) to doing so. We can, however, say from the outset that a reductive view states that the mental is in some sense nothing more than the physical. The hard part is specifying that sense (or senses).

My view can also be rightly called a variety of functionalism, though it is neither the traditional analytic functionalism nor psychofunctionalism: indeed, I combine aspects of the two into my own ‘hybrid functionalism’ in chapter 6.

Reductive/identity views are nothing new in the metaphysics of mental states: in addition to the initial proponents Feigl (1958) and Smart (1959), Lewis (1966; 1980), Armstrong (1968), Kim (1998; 2008), Jackson (2012; Jackson, Pargetter, and Prior 1982), Polger and Shapiro (2016; Shapiro and Polger 2012; Polger 2004), and Morris (2018), among others have embraced one or both labels. And, starting with at least Kim, they have had responses to nonreductive physicalist and sometimes nonphysicalist objections to reductive physicalism (I will discuss the former—and how my own view deals with them—though not the latter, in this work).

But their moves all come at a cost. Lewis (1980), Kim (1998), and Polger (2004) all endorse disjunctive or ‘local’ (Kim 1992; 1998) reductions, such that human pain is one kind of physical state, dog pain another, and Martian pain yet still another. This sacrifices the generality of psychology. And psychology's generality is not optional: it is a necessary part of a theory of, say, pain, to say what all pain states have in common (Block 1978). I call this the *generality objection*. Kim (2008), to his credit, at least addresses the issue: he argues that what they have in common is that they all fall under the concept of 'pain', but this is insufficient: in virtue of what do they all fall under the concept of pain? My view aims to rectify this issue and provide a genuinely reductive view that fully addresses the generality objection.

A consistent theme throughout this work and the development of my view is the invocation of developments in philosophy of the life sciences to help resolve issues in the metaphysics of mind. In particular, I appeal frequently to mechanistic explanation (Bechtel and Richardson 1993/2010; Craver 2007; Machamer, Darden, and Craver 2000) and the Homeostatic Property Cluster theory of natural kinds (Boyd 1991; Griffiths 1997; Wilson, Barker, and Brigandt 2007).

I will of course also address the elephant in the room with every reductive physicalist: multiple realization. If a single mental kind can be realized in or token-identical with different physical kinds, then there is a *prima facie* difficulty that must be overcome to say that the mental kind reduces to a physical kind. For reduction cannot be a relation of one mental kind to many physical kinds (at least not without intermediaries introduced with a principled justification). And if mental kinds are functional kinds—as functionalists such as myself must insist that they are—then multiple realization seems to be a simple and unavoidable fact.

My answer here owes much to philosophers who have addressed the issue before me. From Shapiro (2000; Shapiro and Polger 2012; Polger and Shapiro 2016) I take the idea that genuine multiple realization requires different physical realizers whose physical differences make a difference to how the function is realized. From Kim (1973), Adams (1979) and Polger (2004), I take the idea that mental properties or even states don't need to be realized in the same kind of *stuff* i.e., the same basic substance category to be realized by a single physical category. From Piccinini (2020; 2022) I combine these insights by appealing to physical kinds that share an aspect or subset of their respective sets of properties that underlie their causal powers. 'Metals' are such a kind: the category of metals includes multiple basic substance kinds like tin, brass, aluminum, steel, etc., but it is a single kind nonetheless. I add to Piccinini's analysis by grounding such kinds in the Homeostatic Property Cluster Theory of natural kinds.

My main novel contribution to the multiple realizability issue, however, is to apply an insight from Shoemaker (1981): that functional roles can be multiply realizable *in terms of other functional roles*. I argue that the multiple realizability of these functional roles can be 'offloaded' to a multiple realizability relation between functional roles themselves. The reductionist is then required to show only that the remaining functional roles—those that cannot be further broken down into more 'fine-grained' (Bechtel and Mundale 1999) or lower-level functions—are reducible to physical types.

The work proceeds as follows. In Section One (Chapters 2-4), I articulate and defend a definition of reduction to the physical, and an accompanying definition of physical. Roughly: for mental (or other, e.g., social scientific) kind M to reduce to physical kind P, the two kinds must i) be necessarily nomically coextensive, in terms of all M tokens supervening on and being spatiotemporally coextensive with all P tokens. While identity requires metaphysical necessity, reduction only requires nomic necessity: the former is a purely metaphysical relation, whereas the latter is part metaphysical and part scientific, and science deals in nomic regularities. Further, M must ii) inherit its causal powers from P across all nomologically possible worlds, and iii) the (causally efficacious) properties of P must univocally across all nomologically possible worlds explain the instantiation of the properties of M.

I define ‘physical’ in relation to mechanistic explanation. Mechanistic explanation takes a phenomenon of interest and decomposes it into parts that are delineable both functionally and spatially/structurally/compositionally. It is this ability to be delineated spatially/structurally/compositionally that makes a kind physical. This definition has the advantage of automatically including all physical scientific (i.e., chemical, biological) kinds in the physical, as opposed to only those from microphysics.

In Section Two (Chapters 5 and 6), I switch gears and articulate a novel version of functionalism that I call hybrid functionalism (this is a work about a reductive, *functionalist* theory of mental phenomena, after all). Hybrid functionalism is so named because it combines two extant varieties: analytic functionalism and psychofunctionalism. These two differ in terms of what mental phenomena they allow to be relata of functional (i.e., input, output) relations. Analytic functionalism allows only commonsense or *a priori* psychological kinds, such as folk psychological kinds, plus perception and behavior. Psychofunctionalism adds to this list any (functionally delineated) posit of a true empirical psychology. Hybrid functionalism pictures these two levels as standing in a sort of hierarchical realization relation to each other: psychofunctional phenomena are 'realizers' of states and properties bearing analytic functional roles. Hybrid functionalism then posits that person-level/folk psychological mental kinds such as belief, desire, pain, etc.[1] are i) psychofunctional states or properties that are ii) the core realizers (the part of the realizer unique to that phenomenon) of analytic functional roles. Thus mental phenomena *are* found at the psychofunctional level, but they are what they are in virtue of their relation to analytic functional roles.

In Section Three, I explain how an account of mental kinds can be reductive while answering the multiple realizability and generalization objections. In short: Some mental kinds are directly reducible to physical kinds. The range that are so reducible is greatly increased when we take into consideration i) what is required for something to be a physical kind, especially that kinds can be 'metal' rather than only 'aluminum,' 'copper,' etc., and ii) that for many folk psychological kinds, the actual state itself is the core realizer of the analytic functional role. And while the whole occupant of an analytic functional role's being reducible implies that its core realizer is reducible, the core realizer can be reducible even when embedded in different mechanisms, as in distinct digital video word projectors.

[1] Here I assume that all mental states can be functionally defined. Some mental states seem more amenable to this characterization than others: pains, for instance, seem to have an essentially qualitative component, and it is an open question whether qualitative properties can be reduced to functional roles. I do not further discuss the qualitative issue: my (admittedly more modest) aim is to answer charges raised by nonreductive functionalists, like the multiple realizability objection, and improve on other reductive views by answering the generality objection.

But psychophysical reduction generally can still hold even if not all mental kinds are reducible even with the bag of tools listed above to enhance reduction. For many psychological kinds, *qua* functional kinds, are multiply realizable *in terms of higher resolution functional kinds*. The more we sharpen the resolution, the more we get to a point that any physical difference in the realizer either makes a difference to the realized function—in which case we can keep sharpening the resolution—or it does not, in which case the physical realizers constitute a unified physical kind. And while it is possible for two radically different physical bases to realize the *exact* same physical kind, it is extremely unlikely.

I will here briefly explicate these concepts; more detail will come in the relevant chapters. Any physical difference between two entities is going to make a causal difference. But while all functional differences are causal differences (the way I use function), the reverse isn't true, at least when 'function' is relative to the role of an entity in a mechanistic explanation. To use a hackneyed example, hearts pump blood and make noises, but making noises is not part of the functional role of hearts, and this is true when we consider hearts as mechanistic working parts of a circulatory system as much as it is on teleological views of function (which I do not endorse). When the physical differences between two objects do not make a functional difference, the two objects are going to belong to a single physical kind, albeit a broader one than substances. As Piccinini (2020) puts it, entities that differ physically but not functionally do not participate in genuine multiple realizability but mere 'variable realizability.'

Consider two analog watches with the same configuration of gears. That is, there is a 1:1 correspondence for each gear in terms of size of gear and size of teeth between watches. Further, the overall organization of the gears is the same. Yet one watch has brass gears and the other aluminum ones. While these watches are made of different substances, technically speaking, they are not multiply realized watches. You can even exchange gears between the two!

What underwrites the fact that the two watches are of the same type is that the substances composing them are also part of a single kind—metals. To say that metal is not a valid natural kind simply because it is a broader kind than individual elements which are metals seems absurd. It would be like saying species, but not genuses, families, and so on are natural kinds in evolutionary biology. Shapiro (2000) argues that these kinds need not match our pre-existing kind predicates, either. Using the example of corkscrews like I use the example of watches, he argues that a metal and plastic corkscrew of the same configuration are mere variable realizability. He posits that there is an abstract kind of ‘solids of x rigidity’ which is useful for theoretical-scientific explanations. I will further argue that it is a *bona fide* natural kind, even if it is *post hoc*. While I will hold off on the details until later, what I will say now is that these kinds perform the job of theoretical posits/kind terms in science: they are projectible kinds and support multiple generalizations among their members (Griffiths 1997).

There is a conceptual connection between kinds that are composed of variably, but not multiply realized substances and the notion of ‘subset’ (Wilson 2006; see also Yablo 1992) or ‘aspect’ (Piccinini 2020) realization. The connection is that the members of the kind share a subset of their overall physical properties: aluminum and brass share properties that make them both metals, for example.

The next step is to introduce the concept of resolutions of function. A given function can be captured at a ‘higher’ or ‘lower’ resolution. This is not a novel concept: it is at least hinted at, if not formally defined, by Bechtel and Mundale (1999) when they describe the coarseness or fineness of grain of functional and physical descriptions. Low resolution functions are those that Bechtel and Mundale called coarse-grained: I use the example of a word projector, that is, something that has the function of producing words for consumption by human agents. This diverse category includes scribbles of ink on paper, computer monitors, vocal chords, and computer speakers. We can adjust the resolution (resolution is really a spectrum) a bit up, and distinguish between audio and visual word projectors. Within either category, we can up the resolution yet again, and distinguish between analog and digital projectors. Note that ‘digital visual word projectors’ is still a very broad category: it includes computer monitors and televisions of many varieties (CRT—that’s cathode ray tube for the younger folks—LCD, laser projection), as well as any printer that uses a sort of dot matrix format for distributing ink.

Let’s return to the example of digital, visual word projectors. I think that’s high resolution enough. Here’s why. Their parts all include pixels, which we can define spatially, that is, by location.

This may seem *quite* counterintuitive, as pixels made of ink and those made of liquid crystal (as in LCD, or Liquid Crystal Display monitors) are made of quite different stuff.

Even more so, pixels projected by a standard (what we colloquially call) video projector are just bits of light on a screen! Yet being the same type of physical thing doesn't require being the same stuff (Shapiro 2000; Shapiro and Polger 2012; Polger and Shapiro 2016; Piccinini 2020).

Putting this all together, I argue that at the right resolution of functional mental kinds, they are type-identical to physical kinds. They are not type-identical to *neural* kinds, for in some possible world, some being whose mind is made of something entirely distinct from carbon neurons can have a functionally equivalent mind to ours at this very level. Rather, the physical kinds are those with which our neural kinds share a subset of relevant properties that make them both part of a single *bona fide* natural kind.

What I have said so far gives me a psychophysical identity theory. But the work isn't done. The issue is that while there are genuine psychological phenomena that can be type-identified with physical kinds, our folk psychological kinds—the bread and butter of philosophical psychology—are unlikely to be among them. I am thus only in a slightly better position than Kim: while I may be able to get broader psychological categories than 'human pain,' in that I can get 'human pain and all those enough like it at the right functional resolution'—and maybe this can get me dogs and Martians, depending on the details, it's extremely unlikely that it will get me pain as such. There's still Pain-1, Pain-2, etc. each corresponding to a different higher resolution function. So how do we get pain *as such*? That, after all, is Block's challenge to the reductive theorist.

To meet Block's generality objection, we can appeal to the Homeostatic Property Cluster theory of natural kinds. On this view, natural kinds are categories whose member share a subset of properties in a property cluster. The cluster is homeostatic insofar as the cluster*ing* is reinforced by a causal mechanism: either the same mechanism throughout the kind or distinct mechanisms in multiply realizable functional kinds. As long as the latter disjunct of mechanisms still support the homeostatic clustering of properties such that the kind is projectible (explanations using its referring predicate can subsume generalizations), it is sufficient for natural kind status. I argue that disjuncts of reducible sub-kinds of belief (e.g. human belief, dog belief, Martian belief) will form such a natural kind. This argument is similar to remarks made by Kim (1998; 2008) which I will discuss in detail, but he does not have the conceptual tools to allow these kinds in his ontology while still maintaining the reducibility of the 'sub-kinds;' I do.

Metaphysical Preliminaries

Throughout I discuss various types of mental and physical phenomena. Among these are states, processes, properties, entities, and events. I have seen usages that differ from my own—particularly regarding the notion of 'state'—so I shall first clarify how I understand these terms and their relations. An entity is a discrete particular and an event something that happens to a particular, part of a particular, set of particulars, or spatiotemporal region at or over a time. Events are thus two-place predicates, with the physical specification (particular, set of particulars, spatiotemporal region, etc.) as one place, and time as the other. Processes are a type of event where the particular or spatial aspect changes over the course of the event: it is one way at the commencement and another at the finish. States are also types of events but they are ‘boring events’ (Rey 1997), which is to say that they are the way something is at or for a time, without (salient) change. Properties are attributes of particulars, events, or other properties (or, in the case of relational properties, attributes for which individual tokens are shared by a set of particulars, events, or other properties), which are in turn the ‘bearers’ of properties. Finally, I use ‘phenomenon’ as a catch-all to consist of all of the previously mentioned metaphysical kinds.

I have seen the concepts of ‘state’ and ‘property’ used differently, e.g. Jackson (2012) takes mental states and mental properties to be equivalent. I do not: states are events and properties are attributes of events (or particulars). Beliefs are mental states. They are not predicated of the whole mind or brain, but of a part of it at or for a time (Block and Fodor 1972), such that the time interval will be short for occurrent beliefs, and much longer for ‘standing’ beliefs. They are thus unlike the state of water in my pitcher (liquid) or the state of my computer at the moment, where ‘state of x’ implies that all of ‘x’ is the spatial aspect of the event predicate. Beliefs are not, on my way of speaking, properties. There are corresponding properties, such as ‘being a belief’ that can be predicated of a mental state, or ‘having a belief’ that can be predicated of a person or other believing organism, but these properties are not the same as beliefs themselves. However, events, including states, do *exemplify* properties.

That said, the theory I am proposing is intended to apply to entities, events (both states and processes) as well as properties.

SECTION ONE: PHYSICALISM AND REDUCTION

In this first series of chapters, I want to determine what exactly a reductive physicalism is committed to. For all of the talk about reduction, it's not entirely clear what reductionists are committed to. As Piccinini (2020; 11) acknowledges, it's "a vexed question." Pretty much everyone agrees that for (entity, event, property) A to reduce to B, it must be the case that A is 'nothing over and above' or 'nothing more than' B. But that's not terribly illuminating. At least we can all agree that whatever *else* reduction is, it is a two-place relation.

Putnam (1975a), as well as Antony and Levine (1997), argue that reduction of the mental to the physical is incompatible with the autonomy of the mental, but i) 'autonomy' is just as in need of explication as 'reduction' and moreover ii) some authors (e.g. Bechtel 2007) argue that reduction is compatible with autonomy. Piccinini (2020; 2022) thinks that reduction implies ontological priority of a most fundamental level of reality over other levels; Hemmo and Shenker (2022) deny this. And these are just a few of the disagreements.

Yet coming up with a clear picture of what reduction means is an essential part of articulating a reductive theory: it specifies the burden the theory must meet. In Chapter Two, I will start by defining what reductive physicalism (which I will use interchangeably with 'reduction') is opposed to: nonreductive physicalism. In Chapter Three, I will examine various reductive views in literature—specifically, Lewis's (1966; 1980), Kim's (1992; 1998; 2008), and Polger's (2004; Polger and Shapiro 2016)—and how they define reduction. I will suggest that Kim’s definition of reduction is the most promising, although his own reductive account runs into issues that I will aim to resolve. And in Chapter 4, I will supplement Kim’s definition and provide a positive argument for endorsing it.

Chapter Two / Nonreductive Physicalism

To capture an adequate definition of reduction, I want to ask: what view is reduction in direct opposition with, and why? For the definition of reduction should capture the fundamental clash between reductionists and their closest competitor. I take it that the closest competitor is nonreductive physicalism. While discussing nonreductive physicalism, I will refer to reduction as ‘reductive physicalism’. I will do so because it is extremely important to distinguish the question of what *reductive* physicalism (in opposition to nonreductive physicalism) amounts to from the question of what *physicalism* (including both the reductive and nonreductive varieties, and in opposition to dualist views) amounts to.

Motivations for Nonreductive Physicalism

Antony and Levine, nonreductive physicalists, describe the benefit of nonreductive physicalism as a view that “promised to satisfy our materialist longings without depriving us of our minds. (We were very fond of our minds.)” (Antony and Levine 1997, 83). What do they mean by ‘depriving us of our minds’? They further specify that they want to preserve “both the reality and the causal relevance of the mental” (Ibid).

It is helpful to note that they see Public Enemy #1 (at least among physicalist views) as eliminative materialism of the mental, which they characterize as the view "that the taxonomy implicit in our ordinary intentional psychologizing is incoherent with respect to the taxonomies of established natural sciences" (Ibid 84). The reason issues from the fact that "We are interested in accounting for the fact that mental events are causally efficacious, and … we don't think this can be done—unless it is the case that mental properties possess some systematic connection to physical properties" (Ibid). This quotation helps illuminate why eliminative materialism incenses them to the extent that it does: by positing that there is absolutely no connection between the mental and any scientific taxonomy, it *ipso facto* posits that there is no connection between the mental and the category of phenomena which already have their causal *bona fides.* Eliminative materialism denies the causal efficacy of the mental because it denies the very reality of the mental in the first place.[2]

But they see reductive physicalism as Public Enemy #1A: it is hardly better, if better at all. Per Antony and Levine:

> [R]eductionism is, in our opinion, just another form of eliminativism, since it buys the reality of the mental at the cost of its autonomy … We would not be mollified to hear that mentalistic terms might still be ineliminable; we would not count it a victory if it turns out that mentalistic predicates are simply vulgar specifications of physical

[2] When it comes to eliminative materialism, 'real' is something of a term of art. What it means to be real is to be the referent of a predicate that is used in the best explanations of phenomena in the world. Per Rey (1997) 'sunsets' aren't real in this sense because there is no explanatory posit, 'sunsets' that plays a role in explaining our experiences of sunsets—which we undeniably have, although it seems common to interpret eliminative materialism as committed to the claim that we don't! Sunsets are explained in terms of things like ocular biology, and the physical/chemical laws that dictate how light coming at a certain angle acts in the presence of earth's gravitational pull and the composition of its atmosphere.

> properties, retained to permit communication between the cognoscenti and the science-challenged. No, as realists, we want the mental facts to be both real and different from the physical facts. And thus what we want is a non-reductive materialism (Ibid).

The notion of ‘autonomy’ comes up frequently in discussions of reduction versus nonreductive physicalism (e.g. Putnam 1975a; Fodor 1974; Bechtel 2008). But ‘autonomy’ is at least as fraught as ‘reduction’: Piccinini and Craver (2011) identify nearly a dozen senses of the term, and Shapiro (2017) argues that they missed some. I therefore want to focus on how Antony and Levine specify what they mean by ‘autonomy’ rather than the term itself. They say that they would “not count it a victory if … mentalistic predicates are simply vulgar specifications of physical properties”—what I take them to mean is that mental *phenomena* have their own distinct existence from physical phenomena: mental predicates are not just another way of referring to the same thing. Given their focus on causal efficacy, what they want is for mental phenomena to have causal efficacy *in their own right* and not simply in virtue of being the same as physical phenomena (which, again, have already established their causal *bona fides*).

Commitments of Nonreductive Physicalism

Given Antony and Levine’s comments, it is worth asking what nonreductive physicalism is committed to. This will help clarify what reductive physicalism is itself committed to, via contrast, which will (finally) lead to a proposed definition.

Antony and Levine “want the mental facts to be both real and different from the physical facts.” What I take them to mean is that they want there to be mental states, (other) events, properties and perhaps entities that are not identical (or otherwise reducible) to physical facts.

It is worth asking how this is possible in a way that is consistent with physicalism more generally. For phenomena that are not physical would seem to be nonphysical. The answer is that they are dependent on physical phenomena in a particular way. Specifically, they *supervene* on physical phenomena. As Kim defines supervenience: "Mental properties supervene on physical properties, in that necessarily, for any mental property M, if anything has M at time t, there exists a physical base (or subvenient) property P such that it has P at t, and necessarily anything that has P at a time has M at that time.). Further there is an asymmetric dependence relation: if A supervenes on B, then the way B is determines the way A is, but not vice-versa" (Kim 1998, 9).[3]

This is a plausible description of how nonreductive physicalists might see mental states. Given the multiple realizability of a type of mental state M, it might have various physical bases P, Q, and R. Yet (nomically) the presence of P entails the presence of M, while the converse is not true (since M could be present in virtue of the presence of Q or R).

[3] Kim does not at this point specify the sense of necessity involved in supervenience: "The modal force of necessity involved is a parameter to be fixed to suit one's view of the mind-body relation" (Kim 1998, 10). He will argue later in this cited source that nomic necessity is the relevant type, and I will explain why I agree with him in the next chapter.

Nonreductive Physicalism vs. Property Dualism

But supervenience is too broad to capture nonreductive physicalism. For, as Kim himself points out, mind-body supervenience defines 'minimal physicalism' (Kim 1998, 15), which is an ontological thesis. Yet property dualism, the idea that there exist properties the existence of which *falsifies* (genuine) physicalism, is a type of minimal physicalism in this sense: minimal physicalism is simply the denial of substance dualism such as Cartesian interactionism. And indeed, Antony and Levine's insistence on facts other than physical facts harks to Jackson's (1982) discussion of Mary whose putative learning a new, non-physical fact is supposed to falsify physicalism. Thus, the question of whether nonreductive physicalism is really just property dualism in disguise presents itself.

The key to understanding the difference is to see that the supervenience relationship mind-body supervenience itself "is not [itself] an explanatory theory; it merely states a pattern of property covariation between the mental and the physical and points to the existence of a dependency relation between the two" (Ibid, 14). It does not tell you in virtue of what this pattern/dependency relationship obtains.

For varieties of property dualism, including emergentism, "mind-body supervenience as something that admits no explanation; it is a brute fact that must be accepted with 'natural piety'" (Ibid, 13). Intuitively, this definition makes sense of how Jackson's (1982) conception of *qualia* as distinct from all physical facts differs from Antony and Levine's. For Jackson, there is no physical fact—indeed, no mode of learning other than experience—that Mary can rely on to learn what seeing red is like. That seeing red supervenes on (statistically normal, human) visual experiences of wavelength W in context C is just true: we cannot provide any reason why red, and not purple, supervenes on this wavelength and our visual systems. Indeed, we might go so far as to say that we cannot provide a reason or explanation for why any qualitative state supervenes on these physical parameters at all (Chalmers 1996), though I do not want to commit myself to such a strong claim.

If the basis of the supervenience relationship for a property dualist is brute, then the question arises: what is the grounds of mind-body supervenience for a nonreductive physicalist? Or, for that matter, for a reductive physicalist? I will answer the former question in the next section, and the latter in the next chapter.

Varieties of Nonreductive Physicalism

To examine the nonreductivist's answer to the question of what grounds mental-physical supervenience, it will be helpful to examine some standard nonreductive views. There are three that are relevant to my overall discussion, plus two specific views (Antony's [2003; 2008; 2015; Antony and Levine 1997] and Piccinini's [2020; 2022; ms]. Here, I will discuss the first two general views.

Role functionalism is the view that mental states are higher-order properties (or states/events exemplifying these properties), specifically, the property of having some (first-order) property or state play the causal-functional role characteristic of the mental kind (Levin 2018). I will call these higher-order properties, which specifically quantify over another property (or a state) *playing the role* of such-and-such 'higher-order role properties.' Among those who have explicitly endorsed role functionalism are Antony and Levine (1997), Melnyk (2003), and Loewer (2002). The idea predates these views: Kim (1998, 124n29) believes that Putnam (1969) was the first to introduce the concept of a higher-order role property, and Prior, Pargetter, and Jackson (1982) give a higher-order role property theory of dispositional properties.

Role functionalists posit a *sui generis* relation of 'realization' between mental/functional and physical state—indeed, all nonreductive functionalists posit some sort of 'realization,' their burden is explicating what it amounts to (Polger 2004; Morris 2018). The constraints are that it has to i) be weaker than identity, specifically, for functionalists, it has to ii) allow for multiple realizability and thus support a many-to-one correspondence in the mental to physical 'direction,' yet iii) be strong enough to support supervenience, that is, a determination relation in the physical-to-mental 'direction.' Role functionalists define realization in terms of the role-playing relation. While some (e.g. Polger, Morris, both Ibid) question the definition, it is at least coherent and straightforward.

A major benefit of role functionalism is that it points to a property that is coextensive with all instances of an instantiated functional role. This allows for the possibility that, if mental kind terms are functionally defined, those terms can be rigid designators, since they would refer to the same property (the higher-order one) in all metaphysically possible worlds.[4]

‘Token identity’ was first introduced by Fodor (1974), although Davidson (1970) arguably had a token identity predating Fodor’s introduction of the term; Rey (1997) also endorses it. The third type of view I will discuss, aspect realization, may also be understood as a version of token identity, and I would argue that it is appropriate to read any nonreductive physicalist functionalist not explicitly endorsing role functionalism as a token identity theorist. Token identity is the claim that each token of a mental event (including state) or property is identical to a token physical event or property, but that the mental and physical kind to which a token belongs do not correspond; this holds for any token and mental/physical kind. It does not specify whether or not there is any *degree* of correlation between mental and physical kinds. Thus a burden on a reductive physicalist is to say why some relation short of (metaphysically necessary) type-identity is sufficient for reduction. My own argument—which doubles as an argument for favoring my reductive view over its cousins-cum-competitors—is that nomic equivalence is sufficient but anything short of it is not, and all of the other views fall short of it with respect to basic mental kinds.

Token identity is not completely separable from role functionalism. The way I have defined the latter, the two contain a mutually exclusive commitment: on token identity,

[4] However, if Jackson, Pargetter, and Prior (1982) are right, functional roles are ‘covert’ definite descriptions, that is, shorthand for ‘the state that has such and such a role’ ‘the property of having some property or state play such and such a role.’ In this case, they cannot be rigid designators any more than traditional definite descriptions, such as ‘the all-time home run leader in Major League Baseball.’ I do not wish to rule on this matter, so I will simply note that role functionalism has this *potential* benefit.

functional states *are* on a token basis, their physical realizers, whereas on role functionalism, they are the higher-order properties (or events/states that exemplify them). At the same time, token identity has to appeal to higher-order role properties to define mental *kinds*. That is, while token identity says mental states *are* their physical realizers (realizer is thus identity at the token level), what defines the set of 'beliefs' is the set of physical phenomena that possess the functional role characteristic of belief. Thus *the property of having a belief is* a higher-order role property: the property of having a physical state of some sort play the role characteristic of belief.

I will discuss the third type of nonreductive functionalism—aspect or 'subset' realization—in the next chapter. Suffice to say for now that it is a variety of token identity.

Token identity (including aspect realization) and role functionalism have a common explanation to tell about mental-physical supervenience. The properties of the physical realizers explain the instantiation of the supervening mental phenomena. Because all nonreductive physicalists acknowledge multiple realizability, these explanations will not be univocal for any given mental kind. It is an open question how many distinct realizers—and thus distinct explanations—there are for the instantiation (and thus supervenience) of the tokens of a mental type, one I would expect to vary among mental types. There may only be a handful of realizer types—after all, we only need two for multiple realizability to obtain—or each mental token could have a distinct physical type of realizer. Nonreductive physicalism is fine with either option, so long as each physical token explains the presence of the mental token that supervenes on it.

Chapter Three / What is Reduction?

In this chapter, I will formulate a tentative definition of reduction. To get there, however, it will be helpful to go over some influential views whose authors have granted them the status of reductive or identity theories, making explicit their benefits and shortcomings. The shortcomings, especially, will help in formulating a definition that can succeed where others have failed. I will eventually argue that Kim's definition in his later work (1998; 2008) is the right one, although Kim's own reductive theory, distinct from his definition of reduction, will run into issues that I aim to avoid. I should note from the outset that what I am after is reduction *to the physical*; this criterion is what guarantees the relevance of reduction to mind-body questions, but will also require me to provide a definition of 'physical.'

Early Theory Reduction

Initial talk of reduction (Oppenheim and Putnam 1958; Nagel 1961) involved a relation between scientific theories, or at least predicates within distinct theories. In particular, Nagel (Ibid) thought that reduction involved taking the predicates proprietary to one, 'reducing' theory and connecting them via 'bridge laws' to a 'reduced' or 'base' theory. 'Proprietary predicates' of a theory are something like the kind terms used by scientists of a particular discipline that are partially, if not fully, independent of a theory. 'Cell,' for instance, would be a proprietary predicate of biology: while there were certainly theories of biology that did not posit cells (e.g. Aristotle's), many (in fact, all contemporary) theories of organisms do posit cells, although they may disagree on what precisely constitutes a cell (e.g. whether or not ‘being a cell’ requires ‘having a membrane’ and thus whether or not prokaryotic organisms count as cellular or subcellular).

Bridge laws imply a nomic equivalence relation between the referents of the proprietary predicates of the two theories: wherever there is, say, water, there is H_2O, and this is why water reduces to H_2O. They also imply an asymmetric explanatory relation. To stick with the water example: Water, having the Oxygen atom on one ‘end’ of the molecule and the two Hydrogen atoms on the other ‘end’ (in virtue of the angle of their bonds to the Oxygen atom) ‘pulls’ (really, shifts the probability distribution of) the electrons toward the Oxygen atom, so that the Oxygen end of the molecule has a slightly negative—and the Hydrogen end a slightly positive—charge. This makes the water molecule ‘polar,’ which allows it to more easily mix with other polar substances. Nonpolar molecules, such as most oils, separate from polar substances: this is why oil and water separate (oils are on top because their molecular weight is lighter).

We could thus reduce (say) biological phenomena like cells to biochemical phenomena like macromolecules. Underlying this picture is the idea that reality is carved up into hierarchical levels that correspond to various scientific disciplines: roughly, psychology can reduce to biology, which can reduce to biochemistry, which can reduce to a more basic organic chemistry, which can reduce to physics. That is, there is a ‘unity’ between the sciences, and the reduction relation is what grounds this unity (Oppenheim and Putnam 1958).

Yet this underlying picture is quite problematic. First of all, a given scientific discipline can deal with phenomena that vary greatly in magnitude—physics deals with relations between subatomic particles, but also between planets and stars (Shapiro 2022). Even if we avoid this problem by defining the various levels in terms of some combination of scientific discipline and size (i.e. defining the level below basic chemistry as 'microphysics' rather than 'physics'), there is still the problem that reality is simply not layered in the way that the unity of science picture supposes that it is (Shapiro 2022; see also Craver 2007). Atmospheric chemistry and meteorology deal with planets *as such* exerting gravitational forces on molecules like Oxygen or water (in the case of clouds).

Whether reduction as a relation between theories—hereafter 'theory reduction'—can be sustained in light of these issues is an interesting question, but not one I aim to pursue here. For philosophers of mind who take on the reductive label are generally interested in reduction as a relation between phenomena (states, properties, etc.)—hereafter 'ontological' reduction (Kim 1998; 2008; Morris 2018; Bechtel 2007). And I share this focus; consequently, hereafter, when I say 'reduction' without specifying which of these two kinds, I mean ontological reduction.

Nonetheless, there was something right about the theory reductionists: they saw that reduction involves and requires an explanatory element, which I, like Kim (1998) shall strive to maintain. Further, I will argue—again, agreeing with Kim—that this explanatory relation has to hold between nomically equivalent kinds, thus keeping another key insight of Nagel's bridge law formulation.

Lewis's Contingent Identity

While I am going to directly discuss Lewis's view here, his view is sufficiently similar to Armstrong's (1968) and Jackson's (1998; 2012) that this section can be considered a discussion of their views as well.

Lewis makes the following argument: "[t]he definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states possess the definitive characteristics of experience, they must be the experiences" (Lewis 1966, 17). Lewis later extends this analysis from experiences to other mental phenomena (1972; 1980). By Lewis's own admission, this theory is intended as an identity theory (the title of his 1966 paper is "An Argument for Identity Theory").

Because the identity between mental and physical is contingent (in virtue of what "in fact" is the case rather than the case by "analytic necessity")[5] "the identity is contingent" (Lewis 1980, 125). As Lewis explains it: "I do not say that here we have two states, pain and some neural state, that are contingently identical, identical at this world but different at another. Since I'm serious about identity, we have not two states but one. This one state, this neural state which is pain, is not contingently identical to itself … rather, … the

[5] Given that Lewis wrote this paper prior to Kripke's series of talks that became *Naming and Necessity*, it is unclear if he here means necessity, or analyticity, since we now know that the two come apart. On the one hand, in this passage, he is emphasizing what is true necessarily. On the other hand, Lewis *also* emphasizes that he takes functional roles to be "definitive" of mental kinds, which would make the connection analytic. Perhaps he is best understood as saying that the relationship is—happens to be—both necessary and analytic.

concept and name of pain contingently apply to some neural state at this world" (Lewis 1980, 125).

The contingent nature of the identity is supported by his claim that a state of a hypothetical Martian's "hydraulic mind" which contains "varying amounts of fluid in many inflatable cavities [which] … pervade most of his body" (Lewis 1980, 123), can count as 'pain' as long as it has the same (really, a sufficiently similar) functional role as our pain. Finally, the theory is a variety of functionalism—what McLaughlin (2006) calls 'realizer' functionalism—in that physical properties/events are the sort of mental property/event that they are in virtue of the causal-functional role that they play. Lewis thinks that there is a level of regularity between physical and mental kinds. He takes physical states to occupy a causal-functional role *qua* types or kinds "for a population" (Lewis 1980, 126) where the exemplar for a population is a species.[6] Nonetheless, mental kinds are *bona fide* kinds, which Lewis attempts to account for by distinguishing between what they are and the concept of that state, or the property of having a certain kind of state. In his words, "If the concept of pain is the concept of a state that occupies a certain causal role, then whatever state does occupy that role is pain. If the state of having neurons hooked up in a certain way and firing in a certain pattern [occupies the role] … then that neural state is pain. But the concept of pain is not the concept of that neural

[6] Lewis does not want to require that populations are species. Perhaps two species are sufficiently similar—say a short-beaked and long-beaked variety of the same bird that count as distinct species on what amounts to a technicality) that they count as a population, or perhaps a subset of a species (with e.g., a neurological disorder) is sufficiently different from the species at large to count as its own population. I am thankful to several of my undergraduate students in my Winter 2022 Philosophy of Mind class who wrote insightful papers on this topic.

state" (Ibid 125). Alternately, "We must not identify an experience itself with the attribute that is predicated of somebody by saying that he is having that experience. The former is whatever state it is that occupies a certain definitive causal role; the latter is the attribute of being in whatever state it is that occupies that causal role" (Lewis 1966, 19).

That is, a mental phenomenon (instance of a kind/token of a type) *is* one thing, it is that sort of thing (a member of that kind/type) *in virtue of* something else. This division of labor, so to speak, is grounded in the sense/reference distinction: "The identity theory says that experience- ascriptions have the same reference as certain neural-state-ascriptions: both alike refer to the neural states which are experiences. It does not say that these ascriptions have the same sense" (Ibid).

There is much to be said for Lewis's view. The claim that physical phenomena play causal-functional roles, and that there is no 'functional state' above and beyond physical states was a significant insight, and one of the major motivations I had for turning to reductive physicalism. For it gives a positive argument as to why physical states are the mental states they are: it is because of what they do. I find reductive/identity views that appeal to what Putnam (1967, 435) calls "physical-chemical" states or properties absurd. What about the brain is supposed to make it a mind—its braininess? Why are dead brains not then minds? This insight of Lewis's, while commonplace today, should not be understated for how it advanced reductive views over the first generation of identity theories (Smart 1959; Feigl 1958). Indeed, it is why I start my discussion of (ontological) reductive views with Lewis and not these earlier views.

Second, the distinction between what a phenomenon is, and in virtue of what it is that sort of phenomenon is a crucial distinction that will play a major role in my own view. Lewis couched this distinction in his own descriptive theory of meaning, but we can generalize beyond that. And generalizing beyond that is valuable, since not all theories of meaning agree with Lewis on what constitutes the sense of a predicate or concept. Per Devitt and Sterelny (1999), sense of a natural kind term in a purely causal theory of meaning (i.e., Kripke 1972/1980; Putnam 1975b) is whatever members of the kind share, to be determined by empirical investigation. Yet Lewis is clearly not talking about some underlying ‘essence’ (in the deflationary sense) but something knowable *a priori*.

What he was calling reference roughly amounts to the answer to a ‘what is it?’ question, and what he is calling sense roughly amounts to the answer to a ‘in virtue of what is it the sort of thing that it is?’ question. Elsewhere (Pober 2018), I’ve introduced the distinction between a phenomenon’s physical basis and its constitutive basis. For example: I am an uncle. The physical basis of this uncle (ostensively picking myself out) is my body, but the constitutive basis extends beyond my body to include my brother and his son. For if a sibling of mine did not have a child, I would not be an uncle. The posits of physical basis and constitutive basis also correspond to a ‘what is it?’/‘in virtue of what is it that sort of thing?’ distinction. Because it is sense that is problematic, I will henceforth use ‘constitution’, after my ‘constitutive basis,’ for what Lewis called sense. Because ‘physical basis’ is too vague for our present purposes, I will follow Lewis in using ‘reference’ for answers to the ‘what is it?’ question.

Nonetheless, Lewis's view suffers from two issues (which are not entirely distinct). First, it cannot account for the generality of mental types. The only thing human pain and Martian pain have in common is the functional role. But saying that what unifies a mental kind is its functional role, and not some physical property, is the essence of nonreductive physicalism in the form of token identity. Thus—and this is the second issue—it is not clear how Lewis's view goes beyond token identity. It posits some regularity in physical states that realize a mental kind—which is more than saying the relationship is totally random, which Block and Fodor (1972; see also Fodor 1974) hint at when they appeal to the 'Lashleyan doctrine of equipotentiality' (Ibid). But this regularity exists only within a population. Further, if the main insights of Lewis's view are appropriated by token identity, then it is simply unclear why Lewis's view is preferrable, even if it is distinct.

In fairness to Lewis, his target was not token identity or nonreductive physicalism: the views did not exist when he first articulated his identity theory! Rather, he was participating in a debate between physicalism and various dualisms (Cartesian, emergent), and wanted to give a better argument for physicalism than was on offer at the time. And in his appeal to functional roles, he did so. By the same token, what he demonstrated is that mental phenomena are not ontologically distinct from physical phenomena—but with this the nonreductive physicalist agrees! Rather they are saying that mental kinds, or the properties in virtue of which some token is a member of a mental kind, are not identical to physical kinds or corresponding properties. What is needed to have a Lewisian reductive functionalism is a way of saying mental kinds are what they are in virtue of physical properties that correspond much more broadly with functional roles.

Lewis's failure to distinguish his view from the token identity view is particularly instructive. What his view has in common with those views is that it offers disjunctive explanations across a single mental kind as to why its tokens supervened on the physical states they do. This helps us get at one of the crucial aspects of reduction: it requires that there be a single type of physical realizer for a functional/mental kind, such that the explanation for why the mental tokens supervene on the physical tokens that they do is univocal.

Polger and Identity Theory

Since Kripke's publication of *Naming and Necessity*, it has been widely acknowledged that genuine identity relations carry the weight of metaphysical necessity. Kripke's argument is often given in terms of 'rigid designators' which are names that have the same reference in all possible worlds in which the referent exists. Thus, because 'Hesperus' and 'Phosphorus' refer to the same thing in the actual world, they refer to the same thing in all possible worlds: they both refer to Venus (and Venus is, of course, identical with itself).

Philosophers discussing the possible identity of mental state terms often use the language of rigid designators (Lewis 1980; Jackson, Pargetter, and Prior 1982; Kim 1998; 2008; Polger 2004). But this is technically incorrect: as Soames (2002) demonstrated, natural kind terms can't *be* rigid designators.[7] What they can do is share with rigid designators the property of referring to the same kind of thing across all metaphysically possible worlds: Devitt and Sterelny (1999, 85) call such kind terms "rigid appliers." Thus, if 'gold' is a rigid applier that refers to matter composed of atoms with 79 protons (i.e., matter that has the atomic number 79) in this world, then it refers to this type of matter in all possible worlds in which there exists such matter.

However, not all terms work like rigid designators and natural kind terms. In particular, (definite) descriptions function as what Kripke calls 'reference fixers,' in that, roughly, they act as a mapping function from worlds to items in that world. (Rigid designators are also a sort of function that maps worlds to items, they are simply an uninteresting function in that they map to the same item in every world. To give an analogy, rigid designators are mapping functions in the sense that $f(x) = 5$ is a mathematical function).

Consider Putnam's (1975b) thought experiment regarding Twin Earth. H_2O is water on actual Earth: it has a profile of causal powers that make it 'watery stuff': nourishing human cells, falling from the sky in the right conditions, boiling (i.e. becoming gas) at 100°C at 1 atmosphere of pressure, and so on. Per Kripke and Putnam, 'Water' is a rigid applier and thus refers to H_2O in all worlds where there is H_2O (and lacks a referent in all other worlds). In no world does it refer to XYZ. But 'the stuff that is watery' is not, and thus refers to whatever it is that meets the description of water, and so on and thus refers

[7] Soames's argument, while interesting, is tangential to the current discussion.

to ‘water’ in the real world and XYZ on Twin Earth.[8] Most philosophers advocating some sort of reductive view (Lewis 1980; Kim 1998; Jackson, Pargetter, and Prior 1982) endorse the idea that mental kind terms are nonrigid appliers; all but Lewis explicitly appeal to the fact that functional roles are definite descriptions, and if functionalism is true, then mental kinds are definable in terms of their characteristic functional roles.

Polger, however, is unique in that he accepts and embraces Kripke’s account of identity. His view is thus perhaps the only one that deserves to be called an identity theory.[9] Further, he does not consider his view to be a variety of functionalism. It is a creative and compelling view with much to say in its favor. That it endorses metaphysically necessary type-identity makes its case for being reductive the most straightforward of any view that is not eliminativist.[10] For, as Kim acknowledges: “[t]his is an open-and-shut affair if anything in philosophy ever is: Identities do reduce. For reduction nothing works as magically as identities” (Kim 2008, 100).

However, what Polger gains by embracing type-identity head on comes at a cost. First, his view ultimately fails to answer Block’s generality challenge. Second, by abandoning

[8] It is important that the nonrigid designator is specified as ‘the stuff that is watery’, i.e., as a Russellian definite description. For the divide between rigid and nonrigid designators is, as I explained in chapter 3, not a distinction between underlying and superficial or functional properties, but between kind terms and descriptions (Putnam 1975b).

[9] Polger also has an ingenious argument as to why this identity would not be immediately obvious to us, but it is beyond the scope of the current discussion.

[10] Polger himself eschews the label of ‘reduction’ (see Chapter 6.3 of his [2004]) but he is avoiding the label because he understands reduction to either i) be a theoretical rather than ontological issue and ii) require the irrelevance of proprietary psychological explanation. That he counts as reductive on my/Kim’s conception of the term can be seen in that he sees mechanisms and mechanistic explanation (for ontology and theory, respectively) as a viable alternative that he endorses, whereas I argue these are at least compatible with, if not implying, reduction.

functionalism, it encounters difficulties in answering the ‘sense’ or ‘in virtue of what is that physical phenomenon that kind of mental phenomenon?’ question.

As I said, mental states type-identical to physical states for Polger, with no hedging. This is in part possible because of his rejection of functionalism. It is not that functions play no role in Polger’s theory, rather, it is that they play a significantly different—and reduced—role compared to functionalisms proper. Functionalisms—both the nonreductive variety and the reductive views of Lewis, Armstrong, Jackson, Kim, and myself—are what Polger calls “metaphysical functionalism.” He defines metaphysical functionalism as the claim that “[t]o be a mental state M of system S is to realize functional state (in the strong sense) F relative to S, or to S and its environment” (Polger 2004, 80). The ‘strong sense’ of functionalism is the claim that “the type-identity conditions for” *all* mental phenomena, and not just some “can be specified purely in terms of their mutual interconnections and relation to stimulus conditions and behavior” (Ibid, 78). For role functionalists, this criterion is met by stipulating that functional states are (or are exemplified in terms of) distinct (second-order) properties than their realizing (first-order) physical states. For token identity and reductive functionalists, it is met in the way that Lewis describes: the sense of the mental concept or predicate is given by its functional role: it is in virtue of that role that the physical state in question counts as the sort of mental state it is. And it is this ‘in virtue of’ question that Polger has in mind when he says “to be a mental state … is to realize [a] functional state,” which he clarifies by saying that metaphysical functionalism answers “questions that ask for an explanation of what it is to be a mental state, property, process, event, or entity” (Ibid).

What role do functional roles play in Polger's theory? They play the role of Kripkean 'reference-fixers' in that they tell us how to find the physical states that are the mental states and subsequently drop out of the picture. For Kripke, definite descriptions are reference-fixers, and recall that functional roles are definite descriptions. Thus 'the evening star'—really, 'the bright star/heavenly body that appears in the evening' fixes the reference of Phosphorus and/or Venus, and once the reference is fixed, Hesperus=Phosphorus=Venus in all metaphysically possible worlds in which Venus exists. And this is true (on Kripke's theory of meaning) even when, in some other possible world, Mercury is the heavenly body that appears in the evening.

This move allows Polger's embrace of genuine type-identity. If a reductive functionalist were to embrace type-identity, she must account for the fact that some being in some far-off metaphysically (but not necessarily nomologically!) possible world might have the right functional architecture for mental states despite little if any common physical basis between our brains and the realizer of its mind, Polger can simply reject it. Because he is *defining* mental kinds in virtue of their physical basis, he can simply stipulate that the creature does not have mental states without any danger of incoherence.

There is a danger of implausibility, however. For it really does seem like such a creature should have mental states! Polger has a two-pronged response that depends on the sense in which the creature's (putative) mind is similar to ours. If the creature's functional architecture is similar *and* there are physical commonalities, then Polger argues that the two minds (or mental kinds) are in fact of the same physical type. He calls this argument the "Kim-Adams principle" after work by Kim (1973) and Fred Adams (1979). Adams points out: “in specifying kinds, we must be careful to keep ‘property kinds’ from ‘stuff kinds’ (Adams 1979, 157). By ‘property kinds,’ Adams means any kind defined by the exemplification of a property or set of properties, whereas by ‘stuff kind’ he means our typical taxonomy for types of substances (elements, compounds, etc.). Human brains and hypothetical androids’ ‘positronic networks’ (as Data from Star Trek has; a favorite example of Antony [2003; 2015]) are clearly not the same kind of ‘stuff’ but there is no *a priori* reason they cannot share some properties. Again, this clearly anticipates the subset/aspect view.

What of creatures whose physical realizers share no properties? Here, Polger deploys a distinct response: he questions whether they really form a unified psychological kind in the first place. He distinguishes between creatures having mental states and having mental states “exactly like our own” (Polger 2004; 12) which he calls ‘empathetic’ mental states. But the idea that radically (physically) different creatures have empathetic mental states is, per Polger, counter to our intuitions: “[i]t may have seemed to David Lewis that Martians could have sensations like ours. But it does not seem to me that even avian or piscine sensations are quite like ours, although I suppose I am inclined to believe that birds and fish have sensations of some kind” (Ibid, 14). Polger’s intuitions look quite strong if we add in—as he does—a point made by Bechtel and Mundale (1999), that when considering multiple realizability, we have to look at coarseness or fineness of grain. It is highly unlikely that birds, fish, or Martians have pains like ours *at a very fine grain*, much more unlikely than any of them having pain *simpliciter*.

But this response is giving up the game, as far as my interests go. Polger may be willing to countenance species-specific psychologies like Kim, but I already noted that I see Kim as throwing in the towel. When we are defining mental types at a very fine grain, we are by definition giving up on the possibility of commonalities among folk psychological kinds, which are quite coarse-grained.

Further, Polger's moves fall a bit short of heading off any multiple realizability objection. Surely he is right that it is *unlikely* that a creature will be exactly like us functionally—at a fine grain—and entirely different physically. But the set of possible worlds is vast—infinite—so there's something out there that does. It's possible that a creature in another nomically possible world can have the exact same mental states as ours. And this isn't even getting into metaphysically possible worlds, where the might be minds with no physical realizers at all.

Polger might respond that I am begging the question. Because he is defining mental states in terms of their physical realizers, I am making what on his view is an incoherent claim when I talk about creatures with mental states just like ours yet no physical commonality. And while I grant that I, personally, have trouble conceiving of mental states otherwise, there are other ways to define similarity of mental states. Intentional content would be one. Further, since Polger is quite explicit that he is most interested in conscious states, I could define commonality in terms of phenomenal character, and the argument would go through.

What I take away from Polger's view is that genuine type-identity between mental and physical kinds is implausible. For there are metaphysically possible worlds with no matter that could still have minds. Such a world could consist entirely of Cartesian mental substance (see Schwitzgebel [2019] for a vivid description of such a world), as long as the laws of nature in that world allowed for causal relations between 'bits' of Cartesian substance (the 'bits' would have to be non-spatially defined, as would causation, as Cartesian mental substance has no spatial extension). There is at least *prima facie* no *metaphysical* reason that it cannot, at least no more than there is prohibiting p-zombies (Chalmers 1996). Functionalists, understood as those who define mental states in virtue of their functional roles, may be able to be 'functional state identity theorists' (Block and Fodor 1972)—as Putnam (1967) notes, functionalism is compatible with Cartesian dualism (and by implication an 'idealism' where a world has only Cartesian substance, again, as long as the laws of that world supported the right causal relations between 'bits' of the Cartesian substance).[11]

Fortunately for my purposes, while identity is sufficient for reduction, it may not be necessary. Kim articulates a non-identity version of reduction that I believe holds great promise.

[11] Whether such a functionalist could be a physicalist would depend on whether they take physicalism to be a nomological or metaphysical thesis.

Causal Inheritance, Causal Exclusion and Aspect Realization: Setting the Stage for Kim's View

Kim's view of reduction, and my own, has at its center the notion of causal inheritance. Kim offers what he calls the "Causal Inheritance Principle: If mental property M is realized in a system at t in virtue of physical realization base P. the causal powers of this instance of M are identical with the causal powers of P" (Kim 1992, 17). Causal inheritance is (what I am calling) the relationship between M and P. I will discuss later in the chapter how causal inheritance plays into reduction. For the time being, I want to explicate and motivate the notion of causal inheritance itself.

There is a sense in which spatiotemporal causation is localizable, though it is perhaps awkward to speak this way. While Hume notoriously argued that we do not experience causes, merely constant conjunction, when we know that A is the cause of B, say, when a pool cue is the (proximate) cause of a ball's moving, we can localize 'the cause' with the pool cue. And we can localize the proximate cause of the pool cue's moving with the person holding it, and so on. There's a sense in which the cue 'inherits' its causal powers from the person: this is not the sense of 'inherit' salient for reduction. For that, something like spatiotemporal coextension is required.

To see how some phenomenon can be spatiotemporally coextensive (on either the token or type level) yet *fail* to inherit causal powers of another phenomenon, I want to briefly discuss Kim's 'causal exclusion' argument (Kim 1989; 1995; 1998; see also: McLaughlin 2006; Morris 2018).

It issues from the principle that "If C is sufficient for a later event E, then no event occurring at the same time as C and wholly distinct from it is necessary for E" (Kim 1989, 82). Causal sufficiency is just a subspecies of sufficiency, so applying the principle to causation we get 'If C is causally sufficient for a later event E, then no event occurring at the same time as C and wholly distinct from it is causally necessary for E.' Kim gives the principle in terms of events, but it works for any bearer of causal power: if one takes entities or properties to have causal powers as well as events, then they fall under this principle as well. Call it the principle of Causal Sufficiency. Combine this with the principle of the Causal Closure (or Completeness) of the Physical, which states that all physical events have sufficient physical causes. We then get the following putative issue for mental causation:

> Suppose then that mental event m, occurring at time t, causes physical event p, and let us suppose that this causal relation holds in virtue of the fact that m is an event of mental kind M and p an event of physical kind P. Does p also have a physical cause at t, an event of some physical kind N? To acknowledge mental event m (occurring at t) as a cause of physical event p but deny that p has a physical cause at t would be a clear violation of the causal closure of the physical domain … But to acknowledge that p has also a physical cause, p*, at t is to invite the question: Given that p has a physical cause p*, what causal work is left for m to contribute? (Kim 1998, 37).

The example makes the most sense if we take P to be some bit of behavior, m to be its putative mental cause, and p* to be the physical 'realizer' of m (where 'realization' is at least *prima facie* topic-neutral with respect to allowing for the distinctness of m and p*).[12] Given that p has a sufficient physical cause (causal completeness of physical), m is not necessary for p (causal sufficiency). But then we are left with a seeming dilemma: either p has two causes, i.e. is causally overdetermined, or m is epiphenomenal (at least with respect to p). This issue is especially salient for role functionalism, which putatively identifies mental states with something other than their physical realizers *tout court.* Both of these conclusions are unpalatable, hence, another answer is needed.[13]

Enter causal inheritance. If m inherits its causal powers from p*, then it is no mystery how m has causal powers. It has casual powers because p* has causal powers. In a sense, causal inheritance is identity without the metaphysical baggage Kripke gave it. If m inherits its causal powers from p*, then it is because m is, if not technically identical to p*, then nothing more than p*, and this is the essence of what reduction amounts to.

Note that the nonreductive physicalist cannot salvage the situation by invoking token-identity. For while token-identity between a token of m and a token of p* would allow m to inherit p*'s causal powers, it wouldn't inherit them in the right way. If m and p* are merely token-identical, then we have to ask the question: in virtue of what properties does the m/p* event cause P? Without getting into specifics, we can narrow the options down to two property types, a mental property m_p, and a physical property p_p. These aren't just any mental and physical properties: m_p is what makes the event an event of

[12] It helps to remember here that states are events on my ontology.

[13] Schwitzgebel (personal communication) suggests that one option for the role functionalist is to acknowledge that mental states are epiphenomenal, and come up with a story about why we believe otherwise in the first place. I don't know what to say about this option other than i) yes, it is in the conceptual space of possibilities, and ii) it seems anathema to me. If mental causation isn't a thing, then I'm not sure what's preventing a 'zombie Cartesian dualism' from resurging.

type m, and p_p is what makes the event an event of type p (they can thus also be sets of properties, but I abstract away from this detail here). But recall that the causal sufficiency principle applies to properties as well as events. The causal completeness principle, too, applies to properties—indeed, it applies to all phenomena.

As McLaughlin explains: “If token-physicalism is true, then, given Physical Causal Comprehensiveness, mental events have all of their effects in virtue of being instances of physical types. If type-physicalism is false, the question arises whether they also have any of their effects in virtue of being instances of mental types. If they don’t, then it seems that the mental qua mental is causally inert” (McLaughlin 2006, 42). Yet as Kim points out, “antireductionism precludes the … reductive identification [of mental properties] with physical properties” (Kim 1998, 37-8). Bringing back the causal sufficiency principle, the mental properties are epiphenomenal. Thus, while each mental event has causal powers in virtue of being token-identical to a physical event, mental events have no causal powers *in virtue of their being* mental events. Thus the mental is still epiphenomenal.

So far, type-identity is the only solution identified to avoid the causal exclusion problem. Yet nonreductive physicalists have, since the introduction of this argument, come up with an ingenious response. They claim that i) m and p* in some sense overlap without being token-identical or even fully coreferential, and ii) there is a principled reason to choose m over p* as the cause of P.

Yablo (1992) presents an excellent case for how i) is plausible. He argues that mental events and properties (he runs through the argument for each type of phenomenon) stand in the determinable-determinate relation to their physical realizers. Consider the properties of being a square, a rectangle, and a quadrilateral. The first is equivalent to 'having A) four sides B) of equal length, C) connected by right angles,' the second is 'having A) four sides C) connected by right angles,' and the third is 'having A) four sides.' We can thus say a square is defined in terms of properties [A, B, C], a rectangle in terms of [A, C], and a quadrilateral in terms of [A]. Quadrilaterals are determinables, squares are determinates, and rectangles are determinates with respect to quadrilaterals and determinables with respect to squares. Alternatively, we can understand rectangles and quadrilaterals as being defined in terms of a *subset* of the essential properties of a square (and quadrilaterals in particular as being defined in terms of a subset of the essential properties of either squares or rectangles). Indeed, Shoemaker (2001) and Wilson (2011) adapt a similar account of the mental state/realizer relation that they call the Subset accounts. Piccinini (2020; 2022) and Hemmo and Shenker (2022) dub this the *aspect* view: Hemmo and Shenker, in particular, argue that it is compatible with reduction (a topic which I will take up later).[14]

[14] Wilson (2011) and Shoemaker (2001), who use the term 'subset' also both subscribe to the dispositional theory of properties, i.e. that causal powers are conceptually prior to structural/compositional properties. I do not emphasize this point, as Piccinini (2020, 27n14) points out that "The subset view can be reconciled with categoricalism if someone accepts that whether one property realizes another is contingent. This is because, according to categoricalism, which properties confer which casual powers depends on which laws of nature hold. Therefore, which properties confer casual powers that are a subset of those conferred by another property depends on which laws of nature hold." I do think the relation between casual powers and structural/compositional properties is contingent on nomicity, so I accept this reconciliation.

We can thus picture a mental state as [A, C] and its neural realizer as [A, B, C]—the same relation as rectangles have to squares. In the case of mental and neural states, the relevant properties are causally efficacious ones (as long as functionalism is true; if mental states are defined in terms of epiphenomenal *qualia* this principle can be called into question. However, the issue at hand is between reductive and nonreductive varieties of physicalism, specifically functional varieties, so we can rule this possibility out for the time being). Shoemaker and Wilson go further and cash out their Subset accounts in terms of a Causal *Theory* of Properties, i.e., the claim that what it is to be a property is to confer causal powers. While I am sympathetic to this view, we need not go so far: we just need to stipulate that the properties we're discussing are in fact causal(ly efficacious) ones.

There is good reason to think this sort of relation exists. Suppose that at an individual neuron level, my pain state M can be realized in *n* neurons or *r* neurons, where *n* and *r* are sets of neurons, and *r* is missing one of the neurons in *n* but also contains one neuron not in *n*. Assuming that different physical configurations have different causal powers in at least *some* (remote) counterfactual scenario—and this assumption is quite safe, as it follows from the fact that we are discussing causally efficacious properties—then *n* and *r* will differ in terms of their causal properties, but this difference won't make a difference with respect to their roles as realizers of M. Thus the causal powers of M are the subset of causal properties shared by *n* and *r* (or a subset of that subset). This means M cannot be (type or token) identical to either *n* or *r*.

Now suppose that M causes p, and in this specific case, M's physical realizer is *n*. The subset/aspect theorist now concedes to Kim that p can't have two distinct causes, so we need to choose whether M or *n* is most appropriately the cause. Yablo (1992) gives the following example and accompanying story for choosing the cause. Suppose we train a bird to peck at cards of a certain color. And when given cards that are scarlet, yellow, and navy blue, it picks at the scarlet card only. Now scarlet is a determinate of the determinable 'red'. Is the bird trained to pick at red cards or scarlet cards? Put differently, is it the redness of the card or the scarlet-ness of the card that causes the bird to peck at it (ignoring possible issues about misrepresentation). Well, we know that scarlet is also a determinate of 'color', but we can rule out 'color' as the cause of the bird's pecking, since it didn't pick at other colored cards. The way to determine what is causing the bird to peck is to give it cards that are other shades of red and see if it picks at them. If so, it picked at the original card in virtue of its redness; if not, it picked at the original card in virtue of its scarlet-ness.

Woodward (2005) formalizes an account that captures this intuition in his 'interventionist' account of causation (see Woodward 2008; Shapiro 2010 for discussion of the interventionist account of causation and the causal exclusion problem). The most general description of the property which changes the result or effect when altered or removed (intervened upon) is the one rightly called the cause. If no red colors other than scarlet cause the bird to peck at cards, then scarlet is as general as one can go, and thus scarlet-ness is the cause. If some other shades of red cause the bird to peck, then it is redness, or some subset of redness that is wider than scarlet-ness (conversely, if the bird only pecks at some scarlet cards, then it is some particular set of shades of scarlet).

Now return to the case of M and its realizers *n* and *r*. Recall that *n* and *r* have distinct causal powers. If a given instance of M is realized in *n*, but we (somehow) ‘switch out’ *n* for *r*, does that effect the causal powers of M? By definition, it does not, since M is multiply realizable in *n* and *r*. Thus, by the interventionist principle, M is the most general property we can use to describe the cause, and M, not *n* is the proper cause.

I am not at this point interested in whether this response works (for what it’s worth, I think it does, but I also think it is a reductive, rather than nonreductive move; a claim I shall discuss in detail in Chapter 8). The point I want to make is that M does *not* inherit its causal powers from *n* or *r at the type level*, because they have distinct sets of causal powers. Granted, M’s are a subset of *n*’s and *r*’s, but a subset is distinct from the ‘mother’ set: it is a proper parts, and proper parts are (by definition) distinct from their wholes (Wilson 2011). Note that a token of M may inherit causal powers from a realizing token of *n* or *r*. I will discuss later why token causal inheritance is not sufficient for reduction.

On the type or token level, however, causal inheritance is an all-or-nothing affair: two phenomena need to have coextensive causal powers for one to inherit them from the other (this may seem like a counterintuitive use of ‘inherit,’ as a child can inherit part of a parent’s estate: fair enough, but it is a proprietary term of Kim’s, and he defined it that way. Whether my account ultimately works turns on whether or not causal inheritance *as Kim defined it* is at the core of reduction, not on how the term ought to be defined).

Kim’s View: Nomic Equivalence and Local, Functional Reductions

Having defined causal inheritance, I can now explicate Kim’s own reductive account of mental states. I consider Kim to be a reductive functionalist, although he does not use the label himself. I also consider his view to be the closest to my own, though it falls short in a few places. Indeed, my project may be conceived of as augmenting Kim’s view to overcome its limitations.

Kim’s view issues from a certain set of motivations (Polger 2007) that I take seriously as well. First, he wants to overcome the causal exclusion problem, which he introduces in his (1989), and (*pace* Yablo et al) thinks that the only way to do so is to adopt the causal inheritance principle on a token level (Kim 1998); this amounts, for Kim, to a token identity thesis (Kim 2008, 106). But because Kim (1998; 2008) views the project of reduction to be an explanatory one—this is his second concern—he wants more generalizable relations than token identity can provide.

Third, he takes agrees with Fodor that for any mental kind, there are multiple possible physical realizers in the world, or at least other nomically possible worlds (Kim 1992). And finally, while he acknowledges that identity implies reduction (Kim 2008), he also agrees with Kripke (1972/1980) that type identity implies metaphysical necessity, which he does not think can hold between mental/functional and physical kinds. The reason is that if mental and physical properties are to be understood as intrinsic properties, the relationship between them has to be contingent: this is implied by multiple realization. As Kim puts it: “If M[ental] and P[hysical] are both intrinsic properties and the bridge law connecting them is contingent, there is no hope of identifying them. Distinct properties are just distinct, and we can't pretend they are the same" (Kim 1998, 98).

But while identity is sufficient for reduction, Kim does not believe it is necessary. For while Kim is interested in ontological, rather than theory, reduction, he takes some aspects of the Nagelian (1961) model seriously. In particular, he thinks that reduction involves 'bridge laws' connecting two properties, and bridge laws involve i) a biconditional entailment relation with nomic necessity, and ii) an asymmetric explanatory relation. That the first criterion involves merely nomic, and not metaphysical, necessity, is what allows him to craft a reduction relation that is weaker than type identity.

He notes that while intrinsic properties are designated by rigid appliers, *extrinsic* properties are designated by definite descriptions, which are nonrigid. Thus "to reduce a property M to a domain of base properties, we must first 'prime' M for reduction by construing, or reconstruing, it relationally or extrinsically" (Kim 1998, 98). We do this by making it a functional description as functional roles are relational: they are describing what happens when you put a phenomenon with X (intrinsic) causal powers in context C.[15] But the relation between physical and functional can at maximum hold with nomic, and not metaphysical necessity: "M is defined in terms of its causal/nomic relations to other properties, and since these relations are contingent—contingent on the laws that prevail in a given world—it is a contingent fact whether a given property satisfies the causal/nomic specification that is definitive of M" (Ibid, 99). For example, "given the

[15] They thus do not endow causal powers on their own. This point will be important later in this work.

prevailing laws, DNA molecules are the carriers of genetic information in this world, but in worlds with different basic laws, it may well be molecules of another kind that perform this causal work" (Ibid, 100). While this relationship of nomic necessity is not sufficient for genuine type identity, it is sufficient for supporting scientific principles (in fact, given the nature of science, its principles can be no more than nomically necessary). Nonetheless, it is a *vastly* more robust relation than mere token identity, and we can and ought to acknowledge that: Kim therefore says "[a]ccordingly we may say 'M' is nomologically rigid or semi-rigid" (Ibid, 99-100).

Yet multiple realizability implies that mental phenomena do not correspond 1:1 with physical phenomena even with nomic necessity. Rather, like Lewis suggested, the correspondence is relative to a species, or in the case where a single species has multiple realizers, relative to a 'structure' or particular cognitive architecture (Ibid, 94). Kim agrees with Fodor (1974) that identifying a mental type with a disjunction of physical types is problematic in that disjunctive categories are not 'projectible' kinds, where projectibility "is [a] … standard mark of lawlikeness" defined as "the ability to be confirmed by observation of 'positive instances'" (Kim 1992, 11). This definition is equivalent to subsuming generalizations across nomically similar worlds—instances of the kind produce reliable effects as long as the laws of nature are held fixed. Kim thus calls kinds that possess projectability 'nomic kinds' (Ibid, 12). (In a parallel development in philosophy of science, projectability on the basis of underlying causal similarity is thought to be the hallmark of a natural kind on the Homeostatic Property Cluster view of natural kinds [Boyd 1991; Griffiths 1997], which is currently dominant in philosophy of the life sciences).

Rather than reduce a single mental kind like belief to a disjunction of physical kinds, Kim thinks we ought to divvy up the mental kind, such that each kind of physical realizer matches up with a distinct mental 'sub-kind' (not Kim's term). His logic is: if we differentiate physical kinds by their causal powers, then each realizer in the disjunct has different causal powers and thus so do the mental kinds. As he puts it; "If pain is nomically equivalent to N, the property claimed to be wildly disjunctive and obviously nonnomic, why isn't pain itself equally heterogeneous and nonnomic as a kind?" (Kim 1992, 15).

Kim's strategy is thus to use the functional role characteristic of a mental kind to create species-specific (really, system-specific) 'local' mental kinds that do stand in biconditional entailment, causally inheriting, and explanatory relations to a physical kind. Per Kim: "[i]f the same psychological theory is true of humans, reptiles, and Martians, the psychological kinds posited by that theory must have realizations in human, reptilian, and Martian physiologies. This implies that the theory is locally reducible in three ways … The important moral of MR we need to keep in mind is this: if psychological properties are multiply realized, so is psychology itself" (Ibid, 20).

This is, I think, where Kim starts to take the idea of local reductions too far. I agree that a single mental kind, at least the broad folk psychological kinds that are the bread and butter of philosophy of mind, will have multiple realizers, even on a narrow definition of multiple realization. But to conclude from this that psychology as the study of mental phenomena *as such* must itself be eliminated is throwing in the towel on the very point that my view is designed to avoid. Kim does not see himself as throwing in the towel, however. Rather, he acknowledges that there is no property of a certain sort that corresponds with mental predicates. In particular, there is no property under a 'sparse' conception of properties.

The idea of 'sparse' versus 'abundant' properties harkens back to Lewis (1983), who defined the former as a conception of properties where only intrinsic properties that instantiate causal powers count (and thus underwrite projectable kinds) and the latter as a view where any predicate necessarily has a corresponding property (see also Heil 2003). Functional roles do not count as *bona fide* properties under the sparse conception: they are extrinsic, rather than intrinsic, and describe, rather than confer, causal powers. Kim himself prefers a sparse conception of properties (Kim 1998, 105) but one need not endorse that properties only exist if they exist under a sparse conception. An abundant-conception enthusiast can still see the difference between properties the instantiation of which confers membership in what Kim calls a nomic kind, and functional roles do not.

Kim, however, believes he still has a response to Block's challenge. Per Kim, while only a certain subset of predicates correspond to 'sparse properties,' a larger subset—but a subset nonetheless—correspond to *concepts*. 'Chair' is a concept, 'the set of all chairs, my dog, and my clean laundry' is not. Functional concepts, according to Kim, "specify … a pattern of connection[s]" (Kim 1992, 23). And we can then say that "what all pain instances have in common is merely the fact that they all fall under the concept of pain as given by its functional characterization—no more and no less" (Kim 2008, 109).

There is a great deal to appreciate about Kim's work. The causal exclusion problem has dominated discussions of the metaphysics of mental states ever since Kim introduced it three decades ago. His solution involving causal inheritance, nomic equivalence, and explainability is a basically correct account of ontological reduction.

Further, his arguments for the sufficiency of nomic equivalence rather than genuine identity for the purposes of reduction are nuanced and, in my opinion, *extremely* underappreciated. To draw them out explicitly: reduction is simultaneously a scientific and philosophical project, and scientists are interested in projectability of kinds. But projectability is an affair that cannot be broader than nomicity: it cannot survive a change in the laws of nature. We thus need *nomic* projectable kinds. Indeed, when Kim argues against identification of mental phenomena with disjunctive realizers, he phrases it as: "if we insist on having M as a disjunctive property, we may end up with a property that is largely useless. What good would it do to keep it as a property when it is not a projectible kind that can figure in laws, and cannot serve in causal explanations? … multiply realizable properties are causally and nomologically heterogeneous kinds, and this at bottom is the reason for their inductive unprojectibility and ineligibility as causes" (Kim 1998, 109-10).

As long as we are concerned with projectability, nomic, not metaphysical, necessity is the important kind. Kripke can keep his identity: we still have our reduction.

Nonetheless there are two shortcomings in Kim's view. First, there is the issue of how to define 'physical' so as not to raise causal exclusion problems for biology and chemistry. Kim is quite aware of the problem, "There is a tendency among antireductionist philosophers … to construe the physical domain excessively narrowly … Perhaps the standard micro-macro hierarchical model encourages the idea that the causally closed physical domain includes only the basic particles and their properties and relations. But this is a groundless assumption" (Ibid, 113). But I am not sure he avoids the problem as well as he thinks he does (Antony 2015). I supplement his view by defining 'physical' in terms of mechanistic explanation.

Second, he still does not, to my satisfaction, answer Block's challenge. There are two ways in which he denies the univocality of folk psychological mental kinds. On the one hand, he argues that even when distinct realizers have the same functional role, they must be treated differently simply in virtue of being physical kinds. I will deny this claim based on analysis from Shapiro (2000; Shapiro and Polger 2012; Polger and Shapiro 2016) and Piccinini (2020; 2022). On the other hand, he argues that many realizer types result in different functional roles at a finer grain or 'higher resolution' of describing the functional role than the grain/resolution at which folk psychological kinds are traditionally described. I will attempt to improve on the idea that functional kinds are mere concepts by replacing the sparse conception of properties with the homeostatic property cluster theory of natural kinds, which, I will argue, does the work that Kim wants a theory of properties/kinds to do better than Lewis's sparse conception.

In short, it is (only) by augmenting Kim's metaphysical assumptions with insights from recent developments in philosophy of the life sciences that I can overcome the obstacles where his—excellent—theory stumbled.

Defining Reduction

Kim's account has what I believe are all of the ingredients to define reduction: causal inheritance, nomic equivalence, and explanation. It is worth emphasizing, though, that the causal inheritance itself must be on a nomic type-nomic type level.

Token causal inheritance is insufficient for reduction because it cannot meet the explanatory aspect of reduction. Explanation of X via Y requires, roughly, that one can answer the question "Why is X the way it is?" by citing various properties of Y, and adding "Y's properties *qua* a Y, possibly combined in a specific way, result in the instantiation of X's properties."[16] The issue with token causal inheritance is that while a token of Y will possess properties that ensure the entokening of X's essential properties, they can't be Y's essential properties—they can't be Y's properties *qua* Y. If they were, then the relationship would be stronger than a mere token relation. The answer essentially amounts to "it just happened to turn out that way."

A stronger relationship would be one of within-world multiple realizability. That is, in this world, all X's are coextensive with, and causally inherit their powers from either Y

[16] I am here disagreeing with Kim (2008) who argues that genuine identities are not relations that support explanation, because there is nothing to be explained. For Kim, explanations require a 'gap' between the *explanandum* and the *explanans*, and "If the identity holds, there is here only one thing, not two, and, to push the 'gap' metaphor a bit, at least two distinct items are needed to create a gap" (Ibid, 101). I agree with Kim about 'gappiness' but as I will explain in the next section, I believe that the *way we describe* the relata of the identity (or, as I will argue, nomic equivalence) relation create a gap.

or Z. Some additional details are required. Do Y and Z ensure the entokening of X’s properties in the same way? Suppose, to borrow an example from Shapiro (2000) in a different context (that I will use in its original context later) X’s are cork remover, i.e., a device for removing corks from wine bottles, that Y’s are corkscrews and Z’s are vacuum suction mechanisms, and that corkscrews and suction mechanisms are the only cork removers on Earth. In this case, there are two distinct explanations for the instantiation of X’s essential properties: a screw and a suction device. Alternatively, Y and Z explain X’s properties in the same way—that is, the differences between Y and Z do not matter with respect to X. In that case, I shall argue that they are a single kind (for reasons I discuss in the next chapter), and really an instance of the next sort of causal inheritance I will discuss.

For now, let’s take the disjunctive case; this is what Lewis (1980) endorses. I will leave it an open question for now whether X can really be a univocal kind on these species-specific reductions Even if we can answer the question posed by Block (1978) of ‘what do all x’s have in common?’ we cannot answer the question 'why do all x's have that which they have in common?’ univocally. This is insufficient for reduction: what is needed is for the reducing phenomenon to have its properties in virtue of the same underlying properties as widely as is possible. For this is what separates reductive from nonreductive physicalism.

The next level of generality would be where X and Y are coextensive throughout the actual world. But this is really no different from the previous case. If there is another nomologically possible Z which can instantiate X, then it is just a matter of luck that it happened never to arise in the actual world. That is, there is no difference between multiple realizability across nomologically possible worlds and multiple realizability within this world.

The key in all of these cases is that there is no *lawlike* relation between the properties of Y and the properties of X. But reductions do need to be lawlike: this is the insight that I believe ought to be preserved from the theory reduction days, in particular, Nagel's (1961) conception of 'bridge laws.' Bridge laws were intended to render nomically equivalent theoretical posits from one science, i.e. biology, with another, i.e. chemistry.

We now have enough to formulate a working definition of reduction. Recalling that we are interested in reduction to the physical, we now have:

> Reduction$_{\text{(def 1)}}$ = X type-reduces *qua* nomological type to Y iff i) Y is a physical type, ii) X inherits its causal powers from Y for all X's and Y's in nomologically possible worlds, and iii) descriptions of the properties of Y univocally explain descriptions of the properties of X.

The work of this chapter is, however, not quite complete. I must say what being physical amounts to, as well as characterize an adequate sort of explanation that can meet iii). I aim to do both via the constructs of mechanisms and mechanistic explanation.

Mechanisms and Mechanistic Explanation

In mechanistic explanation: a phenomenon is explained by a decomposition into a lower-level that consists of its parts, their activities, and their organization (Bechtel and

Richardson 1993/2010; Machamer, Darden and Craver 2000). The parts are specified both functionally (by their activities) and spatiotemporally (by their location relative to each other): while investigation can proceed along both the functional and spatiotemporal fronts separately,[17] mechanistic explanation only succeeds when the two match up. The set of activities, parts, and their organization in total is the mechanism; an explanation that invokes it is a mechanistic explanation (Bechtel and Richardson 1993/2010).[18] Shapiro provides an excellent description of how a mechanism works:

> The entities in the bell mechanism include a switch, a battery … an iron core, a steel spring, an iron armature, a screw, a hammer, a bell, and electrical wiring. A description of the activities these entities perform goes like this: The switch is pushed, closing an electrical circuit, causing electricity from the battery to flow through wiring … around the iron core. The resulting electrical field magnetizes the iron core, which pulls the iron armature toward it, thus breaking the contact between the steel spring and the screw, thus breaking the magnetic attraction between the iron core and the armature. The momentum of the armature's movement toward the magnet causes the hammer to strike the bell, after which the spring pulls the armature back to its original position, bringing the spring into contact again with the screw, thus closing the circuit, and starting the process all over again. (Shapiro 2017, 1040)

[17] The spatiotemporal localization of mechanistic parts is relative to other parts. The various parts of your circulatory system are going to occupy a great number of spatially locations throughout your life: mine have been from Seattle to Rome, and from South Africa to the Scottish Highlands! But they are going to be spatiotemporally specifiable in relation to each other.

[18] In contrast to Bechtel and Richardson, Machamer, Darden, and Craver advocate for an 'ontic' conception of mechanistic explanation, where the mechanism and the explanation are one and the same. I have to admit that this puzzles me: explanations traffic in predicates, mechanisms in phenomena, and the two are not the same thing.

The decomposition process—which Bechtel (2007) calls ‘reductive’ in virtue of its appeal to a lower level—is iterative. The parts of one mechanism are phenomena in their own rights, and can be decomposed themselves. Eventually, the decompositions will get down to a level where the entities do not consist of parts (if such a level exists), that of ‘fundamental physics.’ However, the way they get there does not map neatly onto the layer cake model. For a mechanism may decompose into parts of various sciences, such as something sub-cellular, like a virus, interacting with a whole organism, or atmospheric molecules interacting with (the gravitational pull of) a whole planet. Thus the levels are local and relative to the specific mechanism in question (Bechtel 2007; Craver 2007); nonetheless, they can offer a complete global picture since all mechanisms eventually decompose into fundamental physics, though different mechanisms, or even different parts of the same mechanism may take different numbers of steps to get there (Shapiro 2022).

Second, there is the assumption that for each kind predicate in the reducing theory there is a corresponding, extant predicate in the base theory. This issue over-emphasizes the role of theories, and their predicates, and underemphasizes the phenomena they are meant to describe. It is true that there need to be corresponding phenomena at the reducing and base level: it is *not* true that the base level needs to have a ready-made predicate. Rather, the description of the phenomenon at the base level will usually take the form of a set of predicates and the relations between the phenomena they pick out. Indeed, this more complex and messy description is essential for the explanation to proceed.

Simply by explicating mechanistic explanation, I have offered an account of a type of explanation that fits the schema reductive explanations. It explains the properties of the reducing or higher level in virtue of those found at the lower level. That it does so by appealing to components and their organization at the lower level, rather than a single predicate does not make it inappropriate for reduction. Indeed, the type of explanation required by reduction almost always works this way, and it was an oversight on the part of the theory reductionists to fail to notice this fact.

Take the example of reducing ethanol. First of all, it is a compound and its constituents are atomic elements: it is not clear how this is supposed to relate to the traditional layer-cake model as both simple compounds (it is not a macromolecule) and atomic elements are part of chemistry. A description of ethanol at the atomic level would be something like: one Carbon atom, bonded to three Hydrogen atoms and a second Carbon atom; the second Carbon atom is bonded (in addition to the first Carbon atom) to two Hydrogen atoms and an Oxygen atom, the Oxygen atom is bonded (in addition to the second Carbon atom) to a Hydrogen atom. All the Hydrogen atoms only have one bond, specified in the previous description. The description, to be complete, would probably add something about the overall shape of the molecule and how it is constituted by the angles at which the bonds are in relation to each other.

The configuration of these atoms would explain the properties of the ethanol molecule, such as how it binds to specific receptors in the brain that, in sufficient quantity, result in intoxication. If the atomic-level description did not include these details, it could not explain the properties of the ethanol molecule as a whole.

If mechanistic explanation is the explanatory relation in reductive relations, then it also has to be univocal for a reducing kind across all nomologically possible worlds for it to support reduction. For it is this univocality that separates reduction from mere physicalism of the nonreductive variety. We can see this by noting that the mechanistic decomposition relation is a type of realization relation. However, the former, unlike the latter, does not assume that that the phenomenon being decomposed or realized is defined functionally, although in the case of mental phenomena, it is. Additionally, while the realization relation was often assumed to be between two types each designated by a single predicate, for mechanistic explanation it is between that kind of realized type and a realizer type that is expressed in terms of multiple kind predicates and the specific organization of their referents in the mechanism in question.

Mechanisms and the Nature of the Physical

most philosophers discussing reduction—both for and against—saddle it with an unnecessarily narrow definition of ‘physical.’ Specifically, they define physicalism in terms of (the referents of) the predicates used by physics, either ‘microphysics’ or ‘fundamental physics’ or that level plus levels that are directly mereologically composed by it. Here are some examples:

“everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together-roughly, biology is to physics as radio-engineering is to electromagnetism” (Smart 1959, 142).

“Physicalists … widely agree that … physical entities should be characterized by reference to (fundamental) physics” (Wilson 2006, 61).

“The problem of mental causation has thus been trans- formed into one of finding a causal role for mental events in this microdeterministic world, a layered world in which all that happens at higher levels (in particular, the psychological level) is wholly determined by what happens at the micro- physical level” (Kim 1993, 167).

> Higher-level explanations seem threatened in two ways, in a Catch-22-like fashion: damned when fitting in a physical world, and damned if they don’t. On the one hand, when higher-level posits cannot be related to the real furniture of the world, as captured in the laws of macrophysics, they can’t be real things or processes in a causally closed world, can’t really explain anything. On the other hand, if a higher-level explanation can be related to physical processes, it becomes redundant, since the explanatory work can then be done by physics (Schouten and Looren de Jong 2007a, 2).

I could go on! But instead I will note that defining ‘physical’ in terms of the most fundamental physics is problematic. Fodor (1974) stipulated that all sciences other than physics were ‘special sciences.’ But this renders the predicates of biology and chemistry to be not automatically physical. Papineau (2008) considers them *prima facie* nonphysical. An implication of this picture—again, a vestige from theory reduction—is that psychology is just another special science and thus its own ‘higher level’ to be treated the same any other special science. This leads to a generality problem of a different sort (Burge 1995; Antony 2015). This problem suggests that the causal exclusion problem Kim raised can be applied to any special science, and thus the causal relevance of chemicals, cells, and so on is threatened just as much as the causal relevance of beliefs.

Kim (1995) had an ingenious solution to this problem, one that I believe he gave up on too early.[19] He distinguished between two hierarchical ways of going 'above' microphysics: higher levels and higher orders. Higher levels are those standing in mereological relations to each other, and we should assume that some so-called special sciences traffic in phenomena that differ from physics merely mereologically:

> I believe it is crucially important not to construe physical domain too narrowly. The standard micro-macro hierarchical model encourages the idea that the causally closed domain includes only the basic particles and their interactions, this is another groundless assumption associated with the hierarchical picture. Obviously the physical domain must also include gates of basic particles, aggregates of aggregates, and so end; molecules, cells, tables, planets, and biological organisms all belong in the physical system. What then of properties? … Similarly, all micro-based properties, being composed of two hydrogen atoms and one oxygen atom certain chemical bonding, must be considered part of the domain. Otherwise, the system won't be causally closed; one kilogram has causal powers no smaller (Kim 1995, 148-9).

By denying that chemical and biological entities and their properties are *prima facie* nonphysical, Kim can create a principled distinction between them and the nonreductive physicalist's construal of mental properties. For mental properties are 'higher-order' for the nonreductive functionalist: they are the property of having some (first-order) property

[19] In response to criticism, he abandoned this approach in favor of another in his (1998). I will not discuss this latter approach—for a thorough critique of it see Antony (2015)—because I believe the correct solution is a (rather minor!) augmentation of his initial one. He was closer than he thought!

play the right causal role (Antony 2008; Antony and Levine 1997; Melnyk 2003; McLaughlin 2006) and mental states (*qua* events) are the exemplifiers of these properties. Per Kim: "both second-order properties and these first-order realizers are properties of the same entity … a second-order property and its [first-order] realizers are at the same level … Consequently, when we talk of [second-order properties and their first-order] realizers, there is no movement downward, or upward [along] … the micro-macro relation" (Kim 1995, 145).

In other words: functional properties are at the same hierarchical level as their physical realizers, and the competition that leads to causal exclusion only happens when we have this same-level relation. For it is clear that entities and properties at different levels have distinct causal powers and subsume distinct generalizations (Bechtel 2008); they have earned their causal keep, so to speak. But with higher-order properties, we're really talking about two ways of describing the same thing: a property that Φ's, and a property of having a property that Φ's.

For Kim the latter property doesn't even really have its own existence: "[w]e may begin by explicitly recognizing that by existential quantification over a given domain of properties, we do not literally bring into being a new set of properties. That would be sheer magic … By mere logical operations on our notations, we cannot alter our ontology" (Kim 1998, 103). This is why Kim favors some way of reducing functional properties to their physical realizers: if we insist that they are somehow distinct, then they must causally compete, as they seem to be doing the same work. But by acknowledging that they are not distinct, that the higher-order property is just an existentially quantified (“there is some property such that…”) first-order property, and thus just another way of talking about the first-order property, we avoid the dilemma.

But Kim’s move doesn’t quite work. First, it’s not clear that functional properties really are just existentially quantified first-order properties: they are certainly not on the Subset/Aspect view described earlier. And while functional properties are genuinely properties of the same phenomena as their physical realizers, without the specific understanding of functional properties as higher-order, it’s not clear why that should matter.

However, there is a worse problem for Kim’s account. For even if it is the case that functional properties, and not chemical/biological entities and their properties, *qua* mereological sums of physical phenomena are what bring about the causal exclusion worries, biological and chemical phenomena are not safe. As Antony (2015, 4) notes, “many of the proprietary properties of *bona fide* sciences, such as biology, are themselves higher-order, functional properties—think of ‘respiration’ or ‘being a cell’.” She adds to this list: “the first clause of the textbook definition of ‘gene’ is generally ‘the functional unit of heredity.’ Even characterizations that mention DNA generally specify a functional restriction: ‘a length of DNA that codes for a polypeptide” (Ibid, 15).

Thus even if we grant that Kim’s distinction between higher levels and higher orders shows that causal exclusion worries are both present for functional properties and absent for mereological sums of physical phenomena, the properties of chemistry and biology are not safe, and the generalization worry—or enough of it—still applies.

But like I said, I think Kim was close. What he was missing was the notion of mechanistic explanation I raised in the previous section to ground the notion of levels. Recall that mechanisms involve parts that are both functionally and spatially (or compositionally) specified. Entities in chemistry work like this: the function of the whole is a result of what the parts do, such as the Oxygen atom in a water molecule 'pulling' electrons. Biology offers even more vivid examples: cells may be functionally defined, but the functions of a cell are explained in terms of its parts, such as the membrane, nucleus, mitochondria (which I was taught was the 'powerhouse'). Respiration is explained in terms of the activities (functions) of the various parts of the trachea, lungs, as well as their parts.

Mechanistic explanation involves the decomposition of a phenomenon to component parts that can be delineated both functionally and spatially/structurally/compositionally. That is, there has to be a *match* between the functionally delineated parts and the spatially/structurally/compositionally delineated ones. Recall Shapiro's (2017) discussion of a doorbell. A physical part—the switch—corresponds to a functional part: the circuit-closer. Another physical part—the battery—corresponds to another functional part: the energy (or charged ion) storage device. And so on. Per Shapiro, not all purely functional decompositions work this way. Shapiro describes Sternberg's (1969) decomposition of the degraded stimulus recall process. This is the process by which an initial stimulus is shown for a very short amount of time, though enough to enter conscious awareness, and is followed after a short intermission by a second stimulus of the same type (e.g., printed letters). The subjects are asked to determine if the two stimuli are the same (e.g., same letter). Sternberg’s model posited multiple, functionally delineated ‘comparison’ components which did not individually correspond to distinct spatial/structural/compositional parts, or even necessarily to a single such part that performed stimulus-stimulus comparisons (there may be—and for a reductionist will be—some functional decomposition of immediate stimulus memory wherein the functional parts *do* correspond to spatial/structural/compositional parts, but this is orthogonal to my point in mentioning Sternberg’s analysis, which is to provide a contrast to mechanistic explanation).

Take a phenomenon type that is amenable to mechanistic explanation across all of its members, across all nomologically possible worlds. In virtue of the applicability of mechanistic explanation, both the phenomenon itself and the component parts can be said to count as physical kinds. Note, however, that this requirement is weaker than saying that the phenomenon (across all instances/worlds) can be given a *single, unique* mechanistic explanation. Reduction of a phenomenon requires that it has a single mechanistic explanation across all instances and nomologically possible worlds (and this relation only guarantees that the phenomenon is reducible; it doesn’t ‘carry over’ to its parts as well). But reduction is a more demanding notion: being a type of physical phenomenon is a broader category than being a type of reducible phenomenon.

Piccinini (2020) argued that mental phenomena have univocal medium-independent mechanistic explanations. A medium-independent mechanistic explanation delineates parts functionally and assumes these functional parts will correspond to spatial/structural/compositional ones, but (unlike standard mechanistic explanations) not the same spatial/structural/compositional parts across all instances to which the mechanistic explanation applies. Let me illustrate. Cells are, on my definition, a physical kind: thus, 'cell' is a physical predicate. Most terrestrial cells, as far as I know, have a suite of functional components in common that explain their various cellular functions.[20] But it's possible that some cells in some organisms—be they terrestrial, extraterrestrial, or in some other nomologically possible world—lack one or more of these parts. For instance, in terrestrial eukaryotic cells, proteins are transported throughout the cell by the endoplasmic reticulum. But it's possible that some cells somewhere are fortunate enough to have all the proteins travel to where they need to by diffusion, and thus lack (likely by never having evolved) an endoplasmic reticulum. In these cells, the function of 'protein transport' is going to be assigned to the undifferentiated plasma, rather than its own organelle. This means that any mechanistic explanation which appeals to 'protein transport' as a functional component will be multiply realizable in terms of complete or robust (as opposed to medium-independent) mechanistic explanations. But this fact does not render cells unfit to be a physical kind. We can thus propose:

> Physical$_{\text{(def)}}$ = For a kind to be physical, all of its instances across all nomologically possible worlds must be amenable to mechanistic explanation.

We will need to tweak this definition a bit. To see why, consider a worry that arises with this way of understanding 'physical': the possibility of triviality. On the one hand, I have

[20] Properly, cells are the locus of many phenomena of interest that we can put under the label 'cellular functions' (e.g. intercellular communication, mitosis, energy production, etc.)—I will abstract away from these individual functions and for purposes of simplicity, call cells themselves phenomena of interest.

said that the natural kind predicates of biology and chemistry and physical because they fit into mechanistic explanation. Yet if—as I am arguing—mental phenomena are amenable to mechanistic explanation, then the conjunction of these claims implies that mental predicates are physical predicates! Yet if mental predicates are already physical predicates their reduction is much more simple. For a physical predicate to be reducible, it just needs to not be multiply realizable in terms of mechanistic explanations/underlying realizers (which are equivalent). But to argue that the mental is reducible—as I am doing—requires more, specifically, establishing its relationship with physical phenomena in the first place. In other words, we can think of the issue of reduction as turning on two distinct issues: the relationship between some type of phenomena whose referring predicates are 'topic-neutral' and physical predicates, and the relationship within levels of physical predicates (see Papineau 2008 for a similar distinction). The way I have defined physical eliminates the first worry, which ought not be eliminated analytically.

To resolve this issue, we can appeal to Shapiro's (2000) distinction between fully and partially functionally defined properties. Shapiro talks of corkscrews and corked-bottle removers, but notes that only the latter category is fully functional. Cork*screw* species a function, but it also specifies a particular physical way of performing that function: via screw. Whereas mental kinds are functional by definition (Lewis 1966) and 'topic-neutral' in that they do not directly refer to anything physical (Smart 1959) in the way that corkscrews refers to screws. Many paradigmatic biological kind terms, like 'cell' seem to be of the 'mixed' sort. Yet it is not clear that *all* biological kind terms work this way. Antony (2015, 15) points out that 'gene' is usually defined as "the functional unit of heredity." She grants that it *can* be defined as a 'mixed' physical/functional concept, as in "a length of DNA that codes for a polypeptide" (Ibid). But it doesn't seem to be part of the definition of 'gene' that it has to be made of DNA. All viruses contain genetic code for polypeptides, but some only contain and use RNA for this purpose. And while it's not clear that we ought to call such RNA (or DNA) sequences in viruses 'genes,' it is clear that some other organism that clearly has genes could potentially have them coded only in RNA, or something else, elsewhere in this universe or in other nomologically possible worlds. Remember, though, that reduction to a physical kind does not mean reduction to a substance kind. If the set of substances which realize genes *qua* polypeptide coders have enough in common (i.e., an aspect according to aspect realization) then they are still some physical kind. DNA and RNA certainly meet this criterion: they have extremely similar chemical structures (DNA stands for deoxyribonucleic acid, and RNA for ribonucleic acid; thus, the only compositional difference is the inclusion of oxygen atoms at a certain location in the structure of the latter and not the former). Further, they both contain only four units that form two 'base pairs.' Given that the only two substances we know of that play this role are DNA and RNA, it seems at least plausible that other polypeptide-coders will have enough in common with DNA and RNA to make 'gene' reducible to a physical kind. We can thus say that genes are polypeptide-coders with a minimal physical specification of what biological 'coding' is in terms of something like base pairs.[21]

[21] Even 'base pairs' is underspecified here. Really, to demonstrate that the kind is physical requires some spatial, structural, or compositional element to the specification of the part. DNA and RNA use four nucleotides each, with a triplet of nucleotides coding for an amino acid (since there are 20 amino acids, two

Given these considerations, we can thus modify our definition of physical:

> Physical$_{(\text{def } 2)}$ = For a kind to be physical, i) all of its instances across all nomologically possible worlds must be amenable to mechanistic explanation, and ii) the definition of the natural kind predicate must include spatial/structural/compositional specifications for at least some mechanistic parts.

In contrast to this definition, mental kinds may be amenable to mechanistic explanation across all instances of a kind in all worlds, but the definitions of mental kind predicates are purely functional and 'topic-neutral.' Thus it is an open question whether their functional decompositions—what Shapiro (2017) and Piccinini and Craver (2011) follow Cummins (1975) in calling 'task analyses'—correspond to mechanistic explanations. To prove that the mental kind is reducible includes (but is not exhausted by) demonstrating that they do.

But now we have a principled difference between the functional definitions of biology and (the nonreductive physicalists' version of) psychology. Further, we have a principled

base pairs would be insufficient, as there would only be 16 (2^4) possible combinations of nucleotides. If we widen the scope of genes to creatures on worlds with >20 amino acids, and even further to creatures made of (e.g.) silicon rather than carbon, the requirement would be that genes require some sort of biologically realized 'digit' along a 'tape' ('strand' in biological jargon) such that some n-tuplet of 'digits' could code for the basic building blocks of the organism's macromolecule proteins. The genuinely structural and spatial specification is then given in terms of digits along a tape, whose composition may vary.

It is important to be clear that the physical components require specification in some combination of spatial, structural, and compositional terms. Otherwise, the notion of physical risks becoming trivial. Eric Schwitzgebel (personal communication), using Shapiro's example of corkscrews worries, "[c]an I say the corkscrew and the vacuum are the same physical type because at the lowest level of decomposition I care about, they are identical." This would be a worry if the notion of physical didn't require specification in spatial/structural/compositional terms. Screws and vacuums have a functional commonality—applying force—but no necessary structural or spatial commonalities (recall that 'spatial' here is defined as spatial relations *between components*).

distinction that puts the functional and purely structural aspects of biology on one side and the nonreductive physicalist's functional/psychological (or social scientific) posits on the other. Insofar as mechanisms are mereological (a very specific and complex mereology [Bechtel 2007]), the biological side of the divide at least somewhat resembles Kim's notion of higher levels.

Assuaging Some Worries

My definition of 'physical,' combined with an understanding of realization as mechanistic decomposition, assuages several worries about reduction that authors have raised. Piccinini (2020, 11) objects to the "ontological fundamental[ity]" of reduction, which implies that the level of (micro-)physics is somehow conceptually prior to all other levels of description or composition. Piccinini, however, distinguishes ontological fundamentality from physical fundamentality: "[s]omething is physically fundamental just in case it has no physical parts" (Ibid, 12). As Piccinini points out, one can endorse the physical fundamentality of physics, which seems trivially true if one accepts physicalism in the first place, but think that some other level—for holists, the whole universe, for example—is the most ontologically fundamental (Ibid). Mechanistic explanation (which Piccinini endorses) only implies physical fundamentality, not ontological fundamentality. Thus if reduction is defined as a relation to the physical, and physical is defined as fitting into the scheme of mechanistic explanation, then reduction does not imply the ontological fundamentality of physics.

I think the basic motivation of ontological fundamentality, aside from perhaps issuing from a conflation with physical fundamentality, is the idea implicit in Oppenheim and Putnam (1958), and made explicit in Fodor (1974), that every science other than physics is a ‘special science.’ Fodor further argues that if reduction is true, then the phenomenon referred to by every natural kind predicate in a special science must be type-identical to a phenomenon referred to by a natural kind predicate of physics (Ibid, 100). But then the other sciences would be superfluous. Fortunately, he concludes, this type-identity (or even something weaker like nomic equivalence) between kinds in the special sciences and physics cannot be true. For special science predicates capture generalizations that are, as Antony (2008, 166) puts it, “invisible” to physics, and that could not be the case if there were equivalent predicates in physics.

Mechanistic explanation shows how different hierarchical or mereological levels (in addition to not being equivalent to levels of the scientific disciplines [Shapiro 2022]) can capture different generalizations. The nomic equivalence relation does not hold between a phenomenon in biology and a phenomenon at some other level designated by a single predicate, but by the initial phenomenon and one designated by several distinct predicates, i.e., the component parts, and their specific organization at the decomposed level. Thus ‘ethanol molecule’ captures generalizations that Carbon, Hydrogen, and Oxygen molecules, plus the posit of atomic bonds, cannot.

Chapter 4 / What Matters in Reduction

In the previous chapter, I provided a definition of reduction based on Kim's work, as well as considerations from Lewis and Polger. I fleshed out the notions of the physical and of causal inheritance, and I offered a candidate type of explanation to fulfill that aspect of the definition. Here, I want to briefly argue for my definition of reduction before changing gears and discussing functionalism. In particular, I want to take a step back and examine a conceptually prior question: how *ought* we go about defining reduction?

Schwitzgebel's (forthcoming) "pragmatic metaphysics" is a good starting point. Per Schwitzgebel, pragmatic metaphysics' is a "metaphilosophically pragmatic" approach that "rel[ies] on pragmatic criteria to choose among competing metaphysical approaches" (Ibid, 351). By situations where there are 'competing metaphysical approaches,' Schwitzgebel means to pick out situations where we have a (usually) technical and/or proprietary term, and there are several plausible definitions competing for prominence. A reliable (but defeasible) indicator of being in such a situation is when specialists who work on a topic offer competing definitions—just the situation we find ourselves in with respect to reduction. Indeed, as I shall demonstrate, we find ourselves in this situation with respect to reduction along multiple dimensions.

Pragmatic metaphysics says that "metaphysical disputes [which] turn on conceptual or terminological choices … should be responsive to our projects, interests, or values" (Ibid., 353). What I take from Schwitzgebel is that we should define reduction in a way that best captures the motivations that cause people to argue that it is or isn't true. More specifically, we should find the crux of the disagreement between reductionists and antireductionists, and define reduction such that, if it is true, the motivations behind reductive arguments are vindicated

We are thus left with the question: what value debate animates the reduction versus antireduction debate? What motivates nonreductive physicalists to reject reduction, and what motivates reductive physicalists to defend it against these objections? It would also seem like 'What motivated reductive physicalists' in the first place?' should be among these questions, but including it would be a mistake. For the initial reductionists (Feigl 1958; Smart 1959) articulated their view in opposition not to nonreductive physicalism but to dualism (in fact of the emergent variety, but in principle also to Cartesian dualism). It was only after they had articulated their views that nonreductive physicalists entered the picture, promoting what they saw as a way to have our cake and eat it too by endorsing physicalism without the putatively problematic commitments of reductionism.

Kim, I think, hits the nail on the head:

> The shared project of the majority of those who have worked on the mind-body problem over the past few decades has been to find a way of accommodating the mental within a principled physicalist scheme, while at the same time preserving it as something distinctive—that is, without losing what we value, or find special, in our nature as creatures with minds (Kim 1998, 2).

Putnam expands on this point. As he puts it: "The question which troubles laymen, and which has long troubled philosophers … is this: are we made of matter or soul-stuff? To put it as bluntly as possible, are we just material beings, or are we 'something more'? … I will argue as strongly as possible that this whole question rests on false assumptions" (Putnam 1975a, 291). Putnam wants to say that the heart of the traditional question of whether we are matter or 'something more', that is, the debate between dualism and physicalism, is really better captured by the debate between reductive and nonreductive physicalism.

I do not quite agree with Putnam here, but I do believe he is on to something. He rephrases his claim, that "the crux of the matter" behind the dualism/physicalism debate is really:

> that both the Diderots of this world [physicalists] and the Descartes of this world have agreed that if we are matter, then there is a physical explanation for how we behave, disappointing or exciting. I think the traditional dualist says, 'wouldn't it be terrible if we turned out to be just matter, for then there is a physical explanation for everything we do.' And the traditional materialist says 'if we are just matter, then there is a physical explanation for everything we do. Isn't that exciting!' (Putnam 1975a, 295).

Putnam is wrong to think that the traditional debate between physicalists and dualists is just one about whether there is a physical explanation for everything we do. But he is right to say that the issue of physical explanation is important, and I would add *speaks to the same concern* as the traditional physicalist/dualist debate. Putnam equates the two because he thinks that the dualist argues for a soul that is subject to laws, just not physical laws: "[i]f it is built into one's notions of the soul that the soul can do things that violate the laws of physics, then I admit I am stumped" (Ibid, 294). But this is exactly what proponents of 'agent-causal' accounts of free will (e.g. Chisholm 1966) advocate. And he believes that no one arguing for souls takes the issue of reincarnation seriously: after all "Christians believe in resurrection in the flesh, which completely bypasses the need for an immaterial vehicle. So even if one is interested in those questions … even then one doesn't need an immaterial brain or soul-stuff" (Ibid, 295). All I can say to this is that I strongly disagree with the intuition here.

Nonetheless, while the question of whether there is a physical explanation—an explanation in physical vocabulary—for everything we do isn't the same as the issue of whether we have a soul, that the physicalist/dualist and the nonreductive/reductive physicalist clash over different issues, there is *something* about the explanatory question. It is *another way* of getting at the distinctiveness of the mind.

Thus we have to ask: what constitutes there being a mental explanation different from the physical one? Recall that the issue is an explanation *of everything we do*. Putnam here is talking about causal explanations: he shares the presupposition with Descartes that explanations for why we do what we do appeal to internal causes (and indeed, Putnam is as responsible as anyone via his [1963] for rejecting Ryle's denial of this picture). What would it mean for the internal causes—for a physicalist—to be other than physical. I suggest that it is whether the mental phenomena appealed to have causal powers in their own right or whether they merely inherit them from physical phenomena, in the sense of causal inheritance at stake in this chapter. To return to the issue of whether our minds are or are not 'distinctive': nonreductive physicalists have an answer for how they are. They say the mind has causal powers in its own right, different from the casual powers of any physical type. The reductive physicalists deny this claim, and argue we must find some other way to make the mind special (if it is special). If this line of thinking is right, then the question of whether the mental causally inherits its powers from the physical maps perfectly onto what we care about.

Any other issues brought under the heading of whether the mental reduces to the physical are, if I am right, illicitly smuggled in to the question of reduction and do not belong there.

SECTION TWO: FUNCTIONALISM

I noted in the introduction that my account of mental states is a reductive functionalism. While I have discussed what reduction amounts to in the previous section/chapters, and I will augment that discussion and offer my full account in the third section, I now want to change gears and discuss the functionalist aspect of my work. The major motivation for this section, with respect to the overall project, is to fill out my toolkit for responding to the multiple realizability charge. But this section can also be taken as a (shorter) work in its own right. It is even something nonreductive physicalists can endorse!

Chapter 5 / Functionalisms

The topic of this chapter is a distinction between two varieties of functionalism: analytic functionalism and psychofunctionalism. In the next chapter, I will motivate and endorse my own version of functionalism, which is a hybrid of the two varieties discussed here. Before doing so, I want to explicate what the two views of which it is a hybrid amount to—hopefully, in more detail than has been given in extant literature.

However, I think it fair to characterize the current literature as strongly favoring psychofunctionalism. A quite incomplete list of philosophers who I would list as endorsing some version of psychofunctionalism in recent work include Quilty-Dunn in perception (2020), Mandelbaum (2014; 2016; Quilty-Dunn and Mandelbaum 2018), Levy (2016), and Madva (2016) in discussing beliefs and/or implicit bias, Arpaly and Schroeder (2013; Schroeder 2004) in desire, Griffiths and Scarantino (2009; Scarantino and Griffiths 2011; Scarantino 2014), and Pober (2018) in emotion, and Aydede and Fulkerson (2018) in pain. In contrast, analytic functionalism has only a handful of contemporary defenders, including Jackson (2012) and arguably Schwitzgebel (2002; 2013).[22] Most advocates of academic functionalism are found in earlier literature, such as Lewis (1966; 1972; 1980), Armstrong (1968) and Shoemaker (1981).

I believe this bias is a mistake: both analytic functionalism and psychofunctionalism have something to offer, yet neither is complete on its own. Consequently, a major aim of this chapter is to even the score by providing a sustained defense of analytic functionalism as a viable theory for naturalistically oriented philosophers of mind. I make this defense by crafting what I believe is the strongest version from analytic functionalism, which I will elucidate by responding to criticisms of the view. Insofar as psychofunctionalism is contrasted with analytic functionalism, performing these analyses of analytic functionalism will also help make more explicit the commitments of psychofunctionalism.

The distinction between analytic functionalism and psychofunctionalism is about what sort of states[23] can play the roles of inputs and outputs that constitute what it is to be a

[22] Schwitzgebel defines his view as dispositionalist, rather than a version of analytic functionalism. While I will discuss his view in more detail in Section 3, I include along with this submission APPENDIX I, in which I argue that Schwitzgebel's view may be understood as a variety of analytic functionalism.

[23] Or events, processes, or properties. In the majority of cases, the same points apply to a functionally defined state, or event, process, or property. When the entities differ in salient ways, I will explicitly discuss these differences and invoke the relevant kind of entities. Otherwise, I use 'functional state' or

certain kind of mental state. Analytic functionalism is the view that the only candidate types are those available to folk psychology (Block 1978). These include perception (as input), behavior (as output), the propositional attitudes, emotions, and so on.[24] Thus, a belief might be (partially) defined as the state type that, combined with desire, takes perception as input and yields behavior as output.

Analytic functionalism is commonly contrasted with *psychofunctionalism* (Block 1978; see also Rey 1997), wherein any state posited by an empirical psychology (i.e. the sort of theories being proposed in psychology departments) are candidates for relata of functional roles. Here, belief might (partially) be defined in terms of the state that, combined with desire, takes information from multimodal sensory integration systems as input, and delivers information to planning, simulation, and motor systems as output.

Functionalism and Lewis-Ramsey Sentences

On all varieties of functionalism, a mental state is defined in terms of what it characteristically does. That is, it is defined in terms of the types of input it takes in and the type of output it produces. Focusing on the output, we can ask of a process: what is it transmitting? The answer is usually going to be given in terms of some type of

'physical state' as a stand-in for any kind of mental entity that is plausibly functionally or physically defined.

[24] Here I am being intentionally imprecise: discussions of analytic functionalism do not generally provide an exhaustive list or robust principle for what phenomena figure into folk psychology. But there are legitimate questions about its boundary: for instance, whether categories that philosophers, but not 'the folk,' discuss and appear to be at the same 'level' such as aliefs (Gendler 2008) or seemings (Cullison 2010) should so count. Even restricting ourselves to categories extant in common knowledge, it is unclear whether, say, implicit bias is or is not part of folk knowledge. Even if it is now, it plausibly wasn't just a few decades ago. In chapter XX, I will specify further what the relata of analytic functional roles can be.

information (i.e. data about the visual scene for the visual system). Or we can ask of a state: what effects[25] does it have? Different tokens of the same type of state will have different effects across organisms, and even within the same organism at different times. Typically, the belief that the Red Sox won the prior day's game will elicit happiness in me but frustration in my Yankees fan brother (it's a complicated family dynamic). But near the end of the season, if the Red Sox are guaranteed a playoff spot and would get a better matchup if they lost their last game, the belief that they won the game would elicit a more ambivalent, bittersweet emotion rather than the usual positive one. We can also understand this last case as illustrating a third type of functional relation: the effects a state has *given what other states are present.* The belief that the Red Sox won last night has a different effect *in the presence of* the belief that they need to lose to play the lowly Tigers in the playoffs, rather than the formidable Angels. None of these effects are going to define the causal role of that belief: what we are looking for are effects that are *typical* of the kind of state in question, effects that *generalize*. There can be exceptions to these generalizations, but they should be few in number and have a story to tell about why they are exceptions. The set of generalizations that, according to functionalism, defines a mental state type is its *functional role*.

The generalizations constitutive of functional roles can be quite specific, at the level of 'the belief that the government is run by a secret cabal causes the entokening of the belief that Trump won the 2020 election, as well as the belief that coronavirus vaccines are an

[25] Functionalism need not understand inputs and outputs as causal phenomena (Polger 2004; Morris 2018) but I think it is strongest when it does, and I will take functionalism to be causal functionalism unless stated explicitly otherwise throughout this work.

alien plot' or as general as 'beliefs play a role in inferences that form new beliefs whose contents are based on deduction/induction/IBE from the original ones.'[26] When these generalizations are part of common knowledge, as those Lewis (1972) discusses are, I follow Lewis in calling them 'commonsense platitudes.'

Functional roles on both analytic functionalism and psychofunctionalism can be formalized via *Ramsification* (Lewis 1970; 1972), where the name of each state type is replaced with a existentially bound quantifier (which I will often call 'variables'). A given quantifier type is defined in terms of its relation to (some subset of) other variables. So, if a belief that P is identified with the variable *p,* then *p* is, for example [the state that is caused by *a, b,* or *c,* and, in the presence of *l, m,* or, n, produces *x, y,* and *z,* respectively, as output]. A Lewis-Ramsey sentence conjoins segments like these which define each individual variable together into an extremely long sentence that describes a whole psychological theory entirely in terms of the relations variables stand it to each other. Individual variables can then be understood as segments of the full Lewis-Ramsey sentence, which I will call 'partial' Lewis-Ramsey sentences (Lewis [1970, 88] calls these 'definition sentences').

Block deploys the conceptual apparatus of Lewis-Ramsey sentences in his formal definition of analytic functionalism and psychofunctionalism. Per his definition:

[26] As Schwitzgebel (personal communication) has pointed out to me, it may not be the case that mental states individuated very finely are going to have type-individuating conditions suitable for a Lewis-Ramsey sentence. That is, it may be 'beliefs,' or 'beliefs about one's possessions' that are as thin as state kinds with type-identities go. I am sympathetic to this point. But for the purposes of the current discussion I remain neutral on it, as none of the claims I am making turn the variables corresponding to very finely-individuated beliefs.

All functional state identity theories (and functional-property identity theories)[27] can be understood as defining a set of functional states (or functional properties) by means of the Ramsey sentence of a psychological theory … [analytic functionalism identifies mental state S with S's Ramsey functional correlate with respect to a common-sense psychological theory; Psychofunctionalism identifies S with S's Ramsey functional correlate with respect to a scientific psychological theory (Block 1978, 269).[28]

Thus, analytic functionalism is functionalism where the variables in the Lewis-Ramsey sentence are all perceptual inputs, behavioral outputs, or (roughly—see note 21) mental states posited by folk psychology. Psychofunctionalism, on the other hand, takes these types *plus* any process, system, state, property, or event posited by (a true) empirical psychology as candidates for inputs and outputs. Note that psychofunctionalism can include generalizations that are allowed in analytic functionalism: it is simply not restricted to them.

To give a simple example, consider behavior issuing from standard belief/desire psychology. Look at it from the lens of belief (the same exercise could be done from the lens of desire as well). The role that belief plays in its process can be considered an

[27] This difference will be important for the next section, when it comes to articulating the way in which my view is reductive, but can, per footnote 20, be ignored for the time being.

[28] Block uses 'Functionalism' with a capital 'F' to demarcate what I am calling analytic functionalism. I am following convention in calling it analytic functionalism, though I must confess that I am not sure when that particular term came into use. It was present in Shoemaker (1981) and is received enough to be used in the *Stanford Encyclopedia of Philosophy's* entry on 'Functionalism' (Levin 2018). It may be that the term is ill-suited to what I take to be the heart of the view, since it is not analyticity that, for me (*pace* Shoemaker, as we will see in 2.3) determines what makes a role a part of analytic functionalism rather than psychofunctionalism. I am open to alternatives, though I worry that, in using one, the point that I am discussing a particular view that is common in literature could be lost.

aspect of its functional role—a generalization that is part of a partial Lewis/Ramsey sentence. Analytic functionalism might capture this part of the functional role of belief in terms of inputs from perception, and, alongside desires, outputs to behaviors (or intentions, which then output to behaviors). Whereas psychofunctionalism might posit multimodal sensory integration systems between perceptual systems and beliefs, counting *these* systems rather than, say, early vision, as the input to beliefs.[29] And it might posit and motor planning, simulation, and execution systems between beliefs and the muscle contractions that execute and constitute behavior, and thus characterize beliefs as outputting to them, strictly speaking.

Lewis-Ramsey sentences are, most fundamentally, descriptions (Lewis 1970; 1972; Block 1978). And for any object, multiple descriptions can be true of it. I am a human, a primate, an instructor of philosophy, etc. That an object can be divided up in many different ways allows for even more possible descriptions to be true of it After all, many descriptions apply to virtually every phenomenon in existence (I alone am a human, a mammal, a dog owner, a philosophy instructor, etc.). Analytic functionalism and psychofunctionalism are two such descriptions. The claim made by analytic functionalists is that there will be a Lewis-Ramsey sentence true of minds i) wherein the variables all correspond to perceptions, behaviors, or states of the types posited by a folk psychology,

[29] Jake Quilty-Dunn (personal communication) raises an interesting objection to this example, which is that the multimodal sensory integration system is simply a *part of* perception, and thus the example does not distinguish the psychofunctional input from the analytic-functional one. However (as Quilty-Dunn is well aware), we can simply alter the example to make my point. For instance, we can have the intermediate systems between perception and belief be hypothesis formation and hypothesis coherence-checking mechanisms as in Davies et al. (2001). Those who share Quilty-Dunn's worry about my example can replace it with this alternative.

and ii) is nontrivial and useful, in that it (for example) allows us to make accurate predictions about the behaviors of our conspecifics. Psychofunctionalists replace claim i) with the claim i.a) that all the variables correspond to states or processes posited by a true empirical psychology, and keep claim ii), except relative to i.a) rather than i). The debate between the two views is not necessarily that the Lewis-Ramsey sentence of the other view is false of minded beings—a psychofunctionalist may make that argument of analytic functionalism, and vice-versa, but mutual exclusivity of descriptions is not entailed by either view. Rather, the disagreement turns on which description includes variables which best correspond to our mental state kinds.

To give a simple example, consider behavior issuing from standard belief/desire psychology. Look at it from the lens of belief (the same exercise could be done from the lens of desire as well). The role that belief plays in its process can be considered an aspect of its functional role—a generalization that is part of a partial Lewis/Ramsey sentence. Analytic functionalism might capture this part of the functional role of belief in terms of inputs from perception, and, alongside desires, outputs to behaviors (or intentions, which then output to behaviors). Whereas psychofunctionalism might posit multimodal sensory integration systems between perceptual systems and beliefs, counting these systems rather than, say, early vision, as the input to beliefs. And it might posit and motor planning, simulation, and execution systems between beliefs and the muscle contractions that execute and constitute behavior, and thus characterize beliefs as outputting to them, strictly speaking.

The Scope of Functionalism

The sort of functionalism I endorse is that mental kinds are defined in terms of their functional roles. Analytic functionalism and psychofunctionalism are various kinds of stories we can tell about what the relata of those functional roles are. They—and my hybrid functionalism—are compatible with a variety of views along another dimension of functionalism.

It is important to note that, while I define functional roles in terms of causal roles, the two are not the same. The causal role is captured by all of the causal connections a mental kind (via some sufficient number of its members) has to other mental kinds, plus perception and behavior. The functional role *can* encompass the entire causal role, but it *need* not. There are two points to make here. First, functionalist views that do not make the functional role the whole causal role are superior, however, they need a principled way of choosing what aspects of the causal role should be the functional role. Second, there are many different stories one can tell about what that principled way should be, and everything I am saying is compatible, I believe, with any one of them.

If the Lewis-Ramsey sentence is to be read literally as defining mental state types, then its definitions should apply without exception. Holding the extension of a kind term fixed renders the result that any instance of a token state possessing a causal power not specified by the Lewis-Ramsey sentence, or lacking a causal power implied by it, falsifies not only the definition for that type of state, but the *entire* psychological theory captured by the Lewis-Ramsey sentence.

Lewis (1972) rightly says that the definitions captured by a Lewis-Ramsey sentence ought not be understood as exceptionless: defining mental kinds in terms of Lewis-Ramsey sentences should include a "usually" or "most of the time" caveat. This helps matters, but not all that much, since it just means that a mismatch between the extension and the sentence has to be cashed out in terms of some causal powers held/lacked by some members. Indeed, any of the moves I discuss in the next section on behalf of Lewis will ameliorate, but never entirely fix this problem. It is still thus too easy to falsify.

Defining a mental kind in terms of a select few of its causal powers makes extensional mismatches between *all* of the causal powers held by a kind and its specification in the Lewis-Ramsey sentence much less likely to falsify the whole psychological theory. This is especially so if we select causal powers relating one kind to others in 'clusters,' i.e. defining the causal role of desire entirely in terms of its relations to beliefs, other desires, perceptions, and behavior, and defining beliefs in terms of its relations to desires, other beliefs, perceptions and behavior. Rey (1997) calls this 'molecular' functionalism and prefers it for exactly the reason that it is more resistant to falsification than a more 'holistic' functionalism.

All of this is to say that we have good reason to prefer a functionalism where the function of a state kind is defined by part, and not all, of the causal powers characteristic of the kind. The idea of causal-role functionalism, and its varieties (analytic functionalism, psychofunctionalism, hybrid functionalism) tell us what kinds of phenomena can be the relata of a functional role. This other sort of functional view fills in a story about how to determine which among the candidate relata should make the cut.

There are many stories about how we might choose. Cummins's (1975) view, on which the function of a phenomenon is determined by the contributions it makes to the system in which it is embedded (usually, an organism) is a popular one. Newer on the market is Piccinini's (2020) goal-oriented view, in which the function of an agent's mental state or property is determined by goals the agent takes on. However, the most famous is probably the teleosemantic view (Millikan 1984; Dretske 1988) on which the function of any biological phenomenon is determined by its selection history. Unlike the task analysis or goal-oriented views, teleofunctiolism is an historical, rather than 'time-slice' view. There may thus be some complications in integrating it with a standard causal-functionalist view, which is usually understood as a time-slice picture. Thus, though a "Swamp-Man's" 'heart' may have the same causal role as our hearts, standard teleofunctionalism would say it is not a heart because it lacks a selection history, subverting the standard causal role story. I do not explore these issues further as they are tangential to the points I am most concerned with making.[30]

Analytic Functionalism: A Deeper Look

It was David Lewis who first introduced the idea of analytic functionalism in his (1966); he later expanded on it in his (1972) by advocating a specific method for choosing which

[30] Given the power many philosophers assign to the Swamp-Man objection—Schroeder (2004), for instance, raises only this objection to teleofunctionalism and deems it sufficient to abandon the view—I should say that there seems to be an easy way out. Lewis (1980) assigns functional roles relative to a population, such that a 'madman' can have pain if he has a state that has none of the functional properties of pain, is realized in the same kind of physical state as typical pain states in typical members of his population (given as humans in Lewis's paper). I wonder if some similar move is possible for Swamp-Man: his 'heart' is the same physical type of organ as typical hearts in beings who are time-slice sufficiently similar to him (he can, I take it, interbreed with us). This issue is worth pursuing further, but not here.

generalizations about functional roles to include in his Lewis-Ramsey sentence. Specifically, Lewis endorses the following procedure for finding the relevant generalizations: “[t]hink of common-sense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses” (Lewis 1972, 256).

Sydney Shoemaker (1981) has two criticisms of this method. First, he notes that this method will generate false generalizations, which should not be included in a Lewis-Ramsey sentence. Second, and independent of the truth of any generalizations, he argues that commonsense platitudes will be inadequate to determine which generalizations constitute the *meaning* of a mental state term. Shoemaker suggests the analytic functionalist appeal to conceptual analysis in order to find out which generalizations are analytically true of a mental state type and define the terms for those state types via those analytic generalizations. While I agree with Shoemaker’s criticisms—especially the latter—I worry that his proposed alteration to the view fails to solve the very issues he raises. I will first discuss each of Shoemaker’s criticisms, and then his proposed solution. I then turn to a discussion of theories of meaning which will demonstrate that not only is Shoemaker’s method inadequate for determining the meaning of mental state terms, but that the generalizations he takes to be analytic may be false as well!

Shoemaker argues that all things being equal, “we surely do not want to define our mental terms in terms of a false theory” (Shoemaker 1981, 104-5). I take this maxim to also apply to a mostly true *theory* that includes some false *aspects*. And Shoemaker is clearly right that some aspects of the Lewis-Ramsey sentence generated by appeal to commonsense platitudes is incorrect. For many of our commonsense platitudes about mental states are certainly false. For instance, I take it to be a platitude of folk psychology that people revise their beliefs when presented with evidence that contradicts them. However, people nearly universally *double down* on their beliefs and reject the evidence as invalid as a first reaction to its presentation, at least if the belief is about something that matters to them (see Quilty-Dunn and Mandelbaum 2018 for extended discussion).[31,32]

That Lewis's method includes a false platitude is problematic, but to what degree? Rey proposes that it renders Lewis's entire view untenable. According to Rey, "as Lewis states [his] view, should any *one* of these 'commonsense platitudes' turn out to be seriously mistaken, *no* mental terms would apply whatsoever!" (Rey 1997, 186).

Rey is making a point about the nature of Lewis-Ramsey sentences, specifically, the way that theoretical terms are defined at once (in one sentence) and in terms of each other. Going off commonsense platitudes, all sorts of beliefs will be defined as having an output

[31] It might plausibly be argued that the doubling-down is in fact the platitude that is commonly held, and that the intuitive plausibility of rational belief revision as a disposition is something we philosophers hold from taking in too much of the overly-intellectualized views of humans common in philosophy. While I am sympathetic to this criticism in general terms, here, I would suggest that the stereotype about doubling down is better understood as common knowledge associated with a certain personality trait—stubbornness—rather than human nature writ large. While we humans do like to believe that we are more rational than we are, we nonetheless by and large understand that what one is *supposed* to do when presented with evidence contradicting one of our beliefs is to revise that belief.

[32] There are reasons, which I will discuss in the next chapter, for being skeptical that this commonsense-contradicting generalization belongs in the Lewis-Ramsey sentence for analytic functionalism. The key points here are that i) it falsifies a commonsense platitude, and ii) whether it eventually does or does not belong in the Lewis-Ramsey sentence will be determined regardless of its status as common knowledge.

to ‘erase’ themselves when presented with the right input (evidence that contradicts their content). Yet for many beliefs, this won’t happen, so the relevant variables (whether they stand for distinct beliefs, or beliefs as a whole) won’t match the states they are intended to match. But since the Lewis-Ramsy sentence is one whole sentence, the whole sentence will therefore no longer be true of us.

However, Rey is overstating his case. Lewis has two levels of protection from this challenge. First, he does not see the causal relation as providing a necessary connection between cause and effect but a ‘typical’ one (Lewis 1966; 1980). As Putnam (1963) himself has pointed out, this is the paradigmatic relation between relata of causal relations; they are defeasible in certain circumstances. This solution won’t do for platitudes that are *typically* false, though. Fortunately, Lewis’s second layer of protection can handle such cases. He says we should not form our partial Lewis-Ramsey sentences out of “the conjunction of these platitudes” but instead use “a cluster of them—a disjunction of all conjunctions of *most* of them,” as “[t]hat way it will not matter if a few are [typically] wrong” (Lewis 1972, 256).

Thus, the consequences of this issue are not so dire. Still, Shoemaker is right, I think, to say that it would be better to find a method that could avoid placing false generalizations in our theory of mental states, even if their existence in the theory is not fatal.

Shoemaker’s second criticism issues from Lewis’s (1972) claim that analytic functionalism, is supposed to provide the *meaning* of mental state terms. Indeed, that they provide these meanings is part of Lewis’s argument for why his analytic (and reductive) functionalism should be adopted. Per Lewis:

> I do not need to make a case for the [functionalist] identity theory on grounds of economy, since I believe it can and should rest on a stronger foundation. My argument is this: The definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states possess the definitive characteristics of experience, they must be the experiences (Lewis 1966, 17).

Lewis uses the term ‘experience’ rather than ‘mental state’ (though he changes this in his later [1972, 1980] papers). If the “definitive characteristic” of any mental state type is its causal/functional role, then, by hypothesis, the state type is defined in terms of that causal role. And if the type is so defined, the term demarcating it has this definition as its meaning. Shoemaker’s issue is with Lewis’s conception of where we get the details of what these definitive causal roles are.

Shoemaker’s (legitimate) issue with Lewis’s formulation is that, regarding the “’common-sense platitudes’ … while he [Lewis] … suggests that at least some of these have the flavor of analyticity about them, it is by no means clear that he wants to include only those platitudes that could be claimed to be analytic” (Shoemaker 1981, 104). Further, Lewis is clear that when he says commonsense platitudes have “a strong odor of analyticity about them” (Lewis 1972, 257), he does not mean that any given platitude is itself analytic. Rather, what he means is “If the names of mental states are like theoretical terms, they name nothing unless the theory (the cluster of platitudes) is more or less true. Hence it is analytic that *either* pain, etc., do not exist *or* most of our platitudes about them are true” (Ibid). Shoemaker then asks, “if our functional ‘definitions’ are not meant to capture the meaning, or sense, of the mental terms, why should the information included in them be limited to facts that are common knowledge?” (Shoemaker 1981, 105). In other words, if our aim is to give the meaning of mental state terms, and commonsense platitudes *as such* do not give this meaning, why should we worry about whether our generalizations are commonsense platitudes in the first place? Alternatively, if our aim is *not* to give the meaning of mental state terms, then why bother with statements known by most competent speakers of a language.

Shoemaker’s solution is to replace commonsense knowledge as a source for the generalizations to include in the Lewis-Ramsey sentence with “an *a priori* [method], one of conceptual analysis” (Shoemaker 1981, 104). We can then use this method to whittle down all of the possible generalizations we could make about various mental state types to a list that “consist[s] of *analytic or conceptual truths* about the relations of mental states to inputs, outputs and other mental states” (Ibid, italics mine). That is, he takes the generalizations which are analytic—and only those generalizations—to be the ones that give the meaning of a mental state term. It is not that these generalizations are not generally known by competent speakers of English—analytic statements are, or at least can very easily be, commonly known.[33] What the conceptual analysis (presumably performed by philosophers) is doing is determining *which* of those commonly known statements are i) (metaphysically) necessarily true and ii) part of the meaning of a mental state term.

He acknowledges that these analytic generalizations will form only a tiny fraction of the generalizations *true of* mental state kinds. In his own words: "Although the functional property or state picked out by this definition will have … infinitely (perhaps uncountably) many causal features, only a tiny and finite subset of these will be mentioned in the analytic-functional[34] definition. What makes the others essential to the functional property is the fact that they are connected by nomological … [or] metaphysical necessity with those that are mentioned" (Ibid, 110).[35] In this way, he also departs from Lewis, who wants to put the whole set, or something near enough, of commonsense platitudes into the Lewis-Ramsey sentence.

Shoemaker is well aware that, as of his writing, the method of deriving meaning he suggests is controversial, and at odds with the Kripke-Putnam (Kripke 1972/1980; Putnam 1975a) causal theory of meaning.[36] On that view, kind terms are picked out by

[33] I am here assuming something like Kripke's (1972/1980, 39) conception of analyticity, of statements that are both metaphysically necessary and *a priori*. The *a priori* aspect is what ensures their being known by most competent speakers.

[34] When 'analytic' modifies a noun like 'functionalism' or 'functionalist' I use two distinct words. When 'analytic-functional' is a composite adjective modifying a different noun, I use the hyphenated form.

[35] For Shoemaker, whether those other generalizations are connected by metaphysical or nearly nomic necessity depends on the truth of the Causal Theory of Properties. However, I do not want to get into the issue of whether CTP is true or false, at least at this juncture.

[36] There are differences between Kripke and Putnam's views on meaning that I will not discuss here, as they are not relevant for the points I am making. For discussion, see Hacking (2007).

the ‘fixing of a reference,’ often, but not necessarily, via ostension. Kripke illustrates how this works: “imagine a hypothetical (admittedly somewhat artificial) baptism of the substance, we must imagine it picked out as by some such 'definition' as, 'Gold is the substance instantiated by the items over there, or at any rate, by almost all of them'” (Kripke 1972/1980, 135). This ‘definition’ need not include the ostensive ‘over there,’ it could have been, ‘the objects that look and feel like such-and-such, located in the room three doors down, on the left.’

For Kripke and Putnam, once the reference is fixed, then the properties that the kind genuinely possesses—by metaphysical necessity—can be discovered empirically. Any properties or cluster of properties associated *a priori* with a natural kind term are merely *used in* the reference fixation. For all of the potential associations determined this way are potentially false and, even if true, not necessarily so.[37] He illustrates his point with the association of yellowness (property) and gold (natural kind), which, per Kripke, is Kant’s example of an analytic association. First of all, it is in fact false—pure gold is white, and it is impure gold, with a bit of copper in it (which is more common) that is yellow (Putnam 1975b, 170). But let us get around this issue by pretending that ‘gold’ refers to ‘impure gold’ (after all, it might have done so in the original ostensive definition), which is as much a natural kind term as ‘salt water’ is. Even in that case, Kripke argues, “[s]uppose an optical illusion were prevalent, due to peculiar properties of the

[37] There is a sense in which properties putatively associated with a kind through scientific discovery are also only defeasibly associated with that kind (Putnam 1975a, 142), but the sense of defeasibility here is merely epistemic—it is that our scientific ‘discovery’ could have been no real discovery at all but an illusion of one.

atmosphere in South Africa and Russia and certain other areas where gold mines are common. Suppose there were an optical illusion which made the substance appear to be yellow; but, in fact, once the peculiar properties of the atmosphere were removed, we would see that it is actually blue" (Kripke 1972/1980, 118).

Putnam, filling out the other side of the coin, demonstrates that an entire natural kind X can have all the properties associated with a term intended to designate natural kind Y via *a priori* conceptual analysis. His example of 'twin-earth,' where everything is the same except the clear, liquid substance the inhabitants drink and encounter in their oceans is not H_2O but rather 'XYZ.' XYZ has all of the properties associated with water via conceptual analysis, but is not water: H_2O is water. Thus, the properties associated with water via conceptual analysis are neither necessary (per Kripke) nor sufficient (per Putnam) for something's being water.

Once we have discovered the properties actually associated with gold, then and only then can the association between property and natural kind be necessary. These necessary associations are thus *a posteriori*. This is, according to Kripke, the case with having 79 protons in the nucleus of its constituent atoms.

Of the Kripke-Putnam theory of meaning, Shoemaker says, "[i]It is perfectly consistent to accept the Kripke-Putnam account of the semantics of natural kind terms like 'gold' and 'water' while rejecting such an account for mental terms. The notion of a natural kind is not the most luminous of notions; but I do not think we should be bothered if we are required to say that pains, like poisons and mousetraps, are not a natural kind, and lack a scientifically determinable essence" (Shoemaker 1981, 110-11). Rather, "[e]ach mental term has analytically associated with it its functional definition, which gives its meaning … there are many terms in the language for which only such an account is plausible [like poisons and mousetraps], and the analytic functionalist holds that the same is true of words like 'pain' and 'desire'" (Ibid).

Shoemaker further suggests that the Kripke-Putnam theory of meaning is connected with psychofunctionalism in the way that his more Russellian view is connected with analytic functionalism, and that the two varieties of functionalism are simply engaged in different projects (he goes on to give reasons for preferring the analytic-functional project, quite similar to those I will discuss in the next chapter). Per Shoemaker, "[w]e might say that the analytical functionalist looks for functional characterizations that give the "nominal essence" of mental states, while the Psychofunctionalist looks for functional characterizations that give the "real essence" of such states" (Ibid, 105). Nominal and real essences, Lockean terms in origin: are, respectively, the combination of superficial properties proprietary to a kind and that which underlies those properties.[38,39] Shoemaker explains his association of analytic functionalism with a Russellian theory of meaning

[38] In Locke's words: "[b]y this real essence I mean the real constitution of any thing, which is the foundation of all those properties that are combined in, and are constantly found to co-exist with the nominal essence" (Locke 1824: Book III Chapter 6 Section 6).

[39] 'Essence' is a metaphysically loaded term, one which sends some philosophers into conniptions. But I don't mean anything substantive by it, nor, do I believe, does Shoemaker. Rather, we just mean the properties that are i) connected to a kind by metaphysically necessity and ii) associated with the meaning of the term which demarcates the kind. Perhaps, following Wilson, Barker, and Brigandt (2007), we can distinguish this use of 'essence' from what they call 'traditional essences,' which carry the more loaded connotation.

and psychofunctionalism with the Kripke-Putnam view: “[w]hereas on analytical functionalism the infinitely many causal features of pain belong to the property designated by 'pain' in virtue of their nomological [or metaphysical] connections with the causal features which are mentioned in the analytic definition of 'pain', according to Psychofunctionalism these features belong to that property in virtue of their nomological [or metaphysical] connections with a real essence” (Ibid, 111). In other words, Shoemaker thinks that by its very nature (of appealing to empirical findings), psychofunctionalism treats mental state terms like Kripke (1972/1980) and Putnam (1975b) treat natural kind terms like ‘gold,’ ‘water,’ and ‘tiger’—as involving some *underlying* or hidden essence.

The upshot is that Shoemaker sees a dichotomy between an *a priori* method of determining meaning that focuses on superficial properties—where ‘superficial’ corresponds with (objects of) perception, behaviors, and folk psychological mental states—and an empirically sensitive method of determining meaning that focuses on hidden essences.

And this picture is a common one. For Rey, the problem with Shoemaker’s view is that while his method “may be fruitful in providing some constraints on the application of our terms and concepts, at least in the case of explanatorily interesting ones [concepts] … the folk frequently ‘defer to experts.’ They know they don’t know about the nature of things” (Rey 1997, 187). And while Shoemaker is (as noted) not exclusively depending on ‘the folk,’ he is, by making the generalizations that on his view constitute the meaning of a mental state term analytic, things that are known by ‘the folk.’

Rey here is emphasizing an aspect of the Kripke-Putnam view (itself emphasized more by Putnam than Kripke). If some of the properties genuinely associated with a term by metaphysical necessity must be discovered, then it stands to reason that only those well versed in the appropriate methods for discovery (e.g. chemistry for chemical terms) will know what those properties are. Putnam (1975b) and Burge (1979) call this *linguistic division of labor*. According to this doctrine, competent speakers need only know some aspects of the meaning of a kind term. Roughly, they need to know enough to identify and distinguish the kind's members from nonmembers (see Devitt [2006] for discussion of what exactly speakers must know in order to use a term in a language as a competent speaker of that language). We common folk merely 'borrow' the ability to use that term from the experts.

What is important here isn't whether Putnam, Burge, and Rey are right or wrong about the linguistic division of labor applying to mental state terms (I think he is correct). Rather, what is important is that Rey in particular is, in a sense, *agreeing with* Shoemaker. He too sees a dichotomy between the *a priori* method of finding the superficial 'nominal essence' about a mental state term, on the one hand, and an empirical method of finding deep or real essences on the other.

But Rey and Shoemaker are both making a mistake when they claim that the properties discovered via empirical inquiry must be some sort of *hidden* essence.[40] As Putnam

[40] In Rey's defense, he does not, by hidden essence, mean 'underlying essence.' He stipulates, "there is nothing in psychofunctionalism that requires the hidden essence to be, so to say, 'vertical,' as they happen to be in the case of substance terms like 'water' or 'gold' where the essence is provided by the underlying internal features of the referent. There might well be hidden, 'horizontal' essences, as is likely in cases like 'battery,' 'capitalist,' 'ecological niche,' or 'spleen,' where the essence is provided by the external relations

remarks, a "misunderstanding [of this view] that should be avoided is the following: to take the account we have developed as implying that the members of the extension of a natural-kind word necessarily have a common hidden structure. It could have turned out that the bits of liquid we call "water" had no important common physical characteristics except the superficial ones. In that case the necessary and sufficient condition for being "water" would have been possession of sufficiently many of the superficial characteristics" (Putnam 1975b, 159).

Suppose Shoemaker is right that the meaning of mental state terms are given by analytic functionalism. It does not follow that the necessary superficial properties will be determinable *a priori*. For once the relation between the Kripke-Putnam method and superficial properties is properly understood, Shoemaker's justification for his method of conceptual analysis bears no weight. This justification turns on both the (in my view, correct) claim that the definitions of mental state terms are given by analytic-functional roles, and the claim—that we can now see is incorrect—that meanings given by superficial properties must be discovered by conceptual analysis. Yet the upshot of Putnam's claim is that the essences of natural kinds which are superficial *also* need to be

the referent bears to other things … at the same explanatory level" (Rey 1997, 189). Thus, by 'hidden' Rey might *just mean* only knowable via empirical inquiry.
However, if Rey really meant this, then the 'hidden' essences might well be part of an analytic functionalism; indeed, that view is precisely what I will advocate. But insofar as Rey is a psychofunctionalist, it seems difficult for him to accept that the functional roles that are the 'essence' of a mental state are at the same level as 'superficial' functional roles that relate to behavior and perception directly. And while these latter sort of roles are not exhaustive of analytic functionalism, they are certainly a part of it. For a theory to be psychofunctionalist, it seems to me that the salient internal functional connections between systems must *underlie* the capacities that a person has that lead her to behave a certain way given certain stimuli.

discovered by empirical inquiry. And there is good reason to think that Putnam is right on this point.

Consider some currently quite salient examples from medicine and immunology. The term ‘COVID-19’ is a natural kind term of immunology/virology. It refers to a disease caused exclusively by the SARS-COV-2 virus, which is also a natural kind term. For these natural kinds, Putnam and Burge’s point about the linguistic division of labor is surely correct. Despite all the knowledge I possess—and despite the fact that I was raised by biomedical scientists—I could not tell you what distinguishes COVID-19 from the flu, or what distinguishes a SARS-COV-2 virus from an influenza virus. I think the latter may have something to do with the coronavirus having a ‘crown-like’ protein structure somewhere on it (since ‘corona’ means ‘crown’ in some Romance languages—see, e.g. the English word ‘coronation’). My immunologist father, however, could tell you much more about what distinguishes the two kinds of virus, and he (or even better: a virologist) could tell you much more about what distinguishes the two illnesses. Indeed, he could even tell you what properties are *constitutive* of being a coronavirus or an influenza virus. I could not.[41] I can nonetheless use the terms as appropriately as he can; I can get the extension of who is sick with a coronavirus and who is sick with influenza just as right as my father. But my ability to use the terms competently without knowledge of their constitutive conditions is parasitic on the existence of people who *do* know those

[41] There is an upper limit, on this picture, to how wrong I can be on this sort of picture. I have to know some basic facts about viruses: e.g. that they infect living tissues. If I said my computer had COVID-19, my being a competent user of the term ‘COVID-19’ would rightly be called into question.

conditions and can, for example, design a test to tell which, if either, type of virus is present in a body.

Consider on the other hand pneumonia. Pneumonia does not refer to a disease in virtue of its being caused by a specific virus or bacterium; pneumonia has many potential causes (SARS-COV-2 being one of them). Pneumonia is distinguished from other pulmonary diseases entirely in virtue of the symptoms it manifests. Yet here, too, those symptoms are better understood by experts than laypeople such as me. Moreover, the symptoms still need to be discovered.

One might reasonably ask: why can't pneumonia, given that it's just a set of symptoms, be arbitrarily stipulated by a doctor, and then passed down through subsequent generations of medical students from their professors? The answer is that pneumonia is not an *arbitrary* set of symptoms. It is a set of symptoms that co-occur, when they occur at all, *regardless of the underlying cause*. And which symptoms co-occur in various syndromes is exactly the sort of thing that medical scientists discover!

Returning to the topic of mental states: suppose Shoemaker is right and the set of generalizations in an analytic-functional role are connected to each other (and others) by metaphysical (or even merely nomological) necessity in virtue of causal relations between those properties. Multiple realizability of analytic-functional states virtually guarantees that those states won't be connected in virtue of an underlying mechanism that they all share—it is precisely with respect to such a mechanism that they are multiply realizable.[42] Then the connections between aspects of the functional role are, like those between the symptoms of pneumonia, exactly the sort of thing that need to be discovered empirically! I therefore conclude that analytic functionalism must use all the tools at our disposal, to find the generalizations constitutive of an analytic-functional role. And one—perhaps the most important—of these tools is empirical discovery as it is prescribed by the Kripke-Putnam theory of meaning.

It may seem counterintuitive, or even contradictory, to have analytic functionalism use generalizations that are empirically discovered. However, this isn't the case. For the crucial difference between analytic and psychofunctionalism isn't about the way in which generalizations were discovered. Rather, as Block (1978) defines psychofunctionalism, what really distinguishes it from analytic functionalism is that *the kind of entity* that can be relata of the functional role relation include things that are *only* discoverable empirically.

Thus, while empirical inquiry is allowed, it is of a different sort than the empirical inquiry involved in the discipline of experimental psychology. The former is about observing patterns in behavior either without worrying about the mechanisms underlying those behaviors, or determining the mechanisms at a superficial level (though, as I will argue in the next section, less superficial than many think), whereas the latter is precisely about figuring out the 'deep' mechanistic basis for behavior.

[42] In the next chapter, I will argue that, for the multiple realizability of analytic-functional states or roles, the possible realizer relata should be understood as psychofunctionally specified states, rather than physical states as it commonly assumed. But that point is orthogonal to the current one: clearly, there is some multiple realizability involved in analytic functionalism.

An example will help. Take the generalization I mentioned earlier: that people double down on a belief about something that matters to them when presented with evidence that contradicts it. One empirical task is simply documenting the existence and robustness of this generalization, and, indeed, it was documented empirically (Festinger, Riecken, and Schachter 1956). Another empirical task is explaining this behavior but using only the sort of 'posits' that people have been using for thousands of years outside of scientific inquiry. In this case, Festinger (1957) suggests the explanation appeals to the claim that being given disconfirming information hurts: it induces a negatively valenced affective state. A deeper, explanation, however, involves positing systems that are *responsible for* making the reception of disconfirming information hurt. Mandelbaum (2019) posits and details a 'psychological immune system' that protects beliefs that matter to us due to the psychological cost of changing them.

The way I am understanding the term, the first two sorts of empirical inquiry are within the scope of analytic functionalism, whereas the third is not. This is because 'hearing (or reading) a statement' and 'refusing to change one's belief'—the input and output in the first empirical study, are the types of things (perception, behavior, albeit covert behavior) that are allowed as candidate inputs and outputs for analytic functionalism. So too is a negatively valenced affective (i.e. emotional) state. But a 'psychological immune system' is not. It is the sort of entity that is only introduced in the context of a more robust scientific explanation.

I should note that Quilty-Dunn and Mandelbaum (2018) define psychofunctionalism differently: they consider all empirically discovered generalizations to be psychofunctional. Thus, they consider any of the above generalizations relating to cognitive dissonance to be available only to a psychofunctionalist and not an analytic functionalist. However, their usage of psychofunctionalism departs significantly from Block's original formulation.

The upshot of this discussion is to show how analytic functionalism makes room for empirical inquiry. The inclusion of empirical inquiry in analytic functionalism is not itself new: Braddon-Mitchell and Jackson have argued for such an inclusion. But their argument allows for empirical inquiry in only a more limited sense. They say:

> Analytical functionalism can give a major role to cognitive science. It is plausibly part of the folk conception of the mind that the boundaries between the various psychological states are important ones, that to be in mental state M is to be in the "important" state that does thus and so. And cognitive science will tell us what the important states are. The point here is like one often made about why a whale is not a fish. It is a mistake to think that the folk conception counts a whale as a fish. According to the folk conception, the boundary between fish and nonfish corresponds to an important boundary, and science tells us the important boundary separates whales from fish. So the folk conception plus what science tells us means that whales are not fish. (Braddon-Mitchell and Jackson 1999, 86-7).

On their view, the folk determine some sort of criteria for fish-hood, and then empirical inquiry determines what meets it. I am arguing for a more expansive role: one where empirical inquiry determines the properties necessarily associated with a mental state kind—and which constitute the meaning of a mental state kind term.

Psychofunctionalism and Chauvinism

Having explicated my view of analytic functionalism and its resources, I now turn to the shortcoming of psychofunctionalism.[43] Block (1978) argues that psychofunctionalism will inevitably fall prey to chauvinism, or rendering the verdict that beings who by any other plausible criteria possess mental states in fact fail to. The intuition behind this worry is straightforward: psychofunctionalism defines functional states in terms of input and output relations to empirically discovered systems. However, it is rarely, if ever, the case that a basic or general kind of mental state, such as belief, will have one and only one mechanistic realization, that is, realization in terms of the sort of systems that experimental psychology is in the business of discovering. *Even when the mechanisms are given a purely functional description*, belief will be multiply realizable in terms of underlying mechanisms (Shoemaker 1981). Yet for the functionalist, if a state type is defined in terms of a functional role, a state lacking that role is not of that type. Note that in this section I am only laying out the charge and some responses made to it: I will prove the charge valid shortly.

Block (1981) provides a powerfully illustrative example. He gives an example of a hypothetical species of Martian that are in many ways, but not psychofunctionally, equivalent to humans in terms of the inferences they make. Suppose both species make the same inferences, and in the same language (say, English). But their cognitive

[43] Block (1978) also makes some criticisms of analytic functionalism. Responding to them is in a sense beyond the scope of this project, since they do not apply to the hybrid functionalism that I will articulate and endorse in the next chapter. However, they are answerable by the analytic functionalist, and I will show how in Appendix 3.

architectures use different ‘strategies:’ “One strategy would be to represent the information in the machine in English, and to formulate a set of inference rules that operate on English sentences. Another strategy would be to formulate a procedure for translating English into an artificial language whose sentences wear their logical forms on their faces” (Block 1981, 6).[44]

The humans and Martians are not psychofunctionally equivalent for two reasons. First, they process inferences in a different representational language or medium: one in English, the other in formal language. But representational language/media are part of an empirical psychology (Fodor 1975). Second, one and not the other involves translation processes in its performance of inferences. Yet as Block himself asks, rhetorically, “[s]hould we conclude that the Martians are not intelligent after all? Obviously not! That would be crude human chauvinism” (Ibid).

It is not as though psychofunctionalists are unaware of this criticism—Block raised it in the very same paper in which he introduced psychofunctionalism in the first place! There are two general strategies of responding.

The first is to acknowledge that psychofunctionalism involves a tradeoff: specifically, to acknowledge that a psychofunctional theory of, (e.g.) belief will cover less than the full range of believers but argue that the payoffs are worth it. It makes belief something like a

[44] We can further “suppose that the Martian and human psychologists agree that Martians and humans differ as if they were the products of a whole series of engineering decisions that differ along the lines illustrated” (Block 1981, 6), though I will focus on the difference in inferential processes.

natural kind, and it links philosophy of mind up with empirical inquiry, as it ought to be. Fodor (1975) and Carruthers (2006) are most explicit about this issue.[45]

I understand where Fodor and Carruthers are coming from: if forced to make the choice between an empirically sensitive theory of belief and a theory that covered creatures crafted in the most far-fetched thought experiments, I would likely choose the former as well. But I do not think we are forced to make this choice. Indeed, hybrid functionalism can be understood as a theory crafted specifically to navigate this particular version of Scylla and Charybdis; I explain how when I articulate the view in Section 10.

More promising is a sort of response that aims to take the chauvinism charge head-on. Rey (1997) provides an especially strong version of this response. He notes that psychofunctionalism does not necessarily imply a specific level of abstraction. Empirical psychology works at multiple levels simultaneously: that of the neuron, that of the neuronal grouping ('small world'), and that of neurocircuits, just to name a few. As Rey puts it: Block's Martians "might differ from humans at *many* different levels of description: … The question is *whether they differ at the level at which a mature psychology will define psychological phenomena*" (Rey 1997, 190, italics in original).

[45] The charge they explicitly address is that the incorporation of empirical work into a philosophical view somehow makes it less than fully philosophical. They both argue that calling the theory a theory of some other sort—they both embrace the label 'theoretical psychology'—is worth it to incorporate empirical work. I am not making the charge that their views aren't philosophical: I find that sort of criticism misguided at best. I am making the more restricted point that their philosophical views are chauvinist, and it is *ceteris paribus* better if chauvinism can be avoided.

Rey is quite right that the way to solve chauvinism is to couch one's functional descriptions at the right level of abstraction. However, I argue that *any functionalist theory that is sufficiently abstract to capture all minded beings will be analytic functionalism.* Block acknowledges that a functional description can capture both the humans' and hypothetical Martians' inferential processes by characterizing them as making inferential transitions from English input to English output. I will argue that this functional description is—and is characteristic of—analytic functional descriptions. Moreover, an analytic functionalist theory is, by definition, constituted by generalizations that are necessarily true of all (or close enough to all, in a sense I will discuss) believers.

This point has been overlooked, I think, in large part because analytic functionalism is often misunderstood in the way I have discussed.

Rey is right that the way for a functionalist to avoid chauvinism is to cash out functional roles at the right level of abstraction. In the previous section, I demonstrated that analytic functionalism (now properly understood) is abstract enough. What I will now argue is that psychofunctionalism *cannot* be abstract enough. Because analytic functionalism captures generalizations that are something like necessarily true of all believers, it avoids chauvinism, which is the problem of a theory rendering too few organisms as believers.

But psychofunctionalism does not have to. I will now argue that, further, psychofunctionalism *cannot* capture generalizations true of all believers without collapsing into analytic functionalism. Recall Block's Martian example. On my understanding of analytic functionalism, it defines 'inference' via a functional description of the inferences both humans and Martians (and all other infer-ers) make. It thus gets the extension right, at least when it comes to Block's Martians. It is only when we look at the underlying mechanisms that humans and Martians start to look different.

But therein lies the problem. Psychofunctionalism *necessarily* takes mechanisms into account: otherwise the putatively psychofunctionalist view is really analytic-functionalist. A functionalist account absent mechanisms is just an account of functional relations between person-level mental states (and perception/behavior): that *just is* analytic functionalism. Take a functionalism that only discusses mechanisms at the most abstract level (that still counts as discussing mechanisms underlying states, and not simply the states themselves). Block's Martian example is itself more abstract than a robust psychofunctional view of beliefs (like Quilty-Dunn and Mandelbaum [2018] or Carruthers [2013]). But it still involved the invocation of an extra system in Martians, the system responsible for translation from English to the formal language (and back).

Take Fodor's (1975) Language of Thought. It had a single, central empirical claim: that the representational medium of thought had the syntax of a spoken language. Suppose humans and Venusians are behaviorally equivalent and that Venusians, like Martians, make all the same inferences as humans. But the only difference between Venusians and humans is that humans have a syntactic LOT whereas Venusians use, say, a connectionist network. I cannot think of a more basic difference than this, which only includes one empirical issue, and an extremely abstract one at that that. But the possibility of Venusians means we can't define mentality in terms of having a Fodorian LOT: they should count as having minds, but this definition would exclude them. But abstract away from the classic computationalist/connectionist distinction, and all we are left with is the relations between states themselves—not the underlying mechanisms— which is analytic functionalism.

A psychofunctionalist might object that the functional role of beliefs might require them to be connected to some information-processing *system* such as an inferential system. And that talk of such systems—and of information-processing as such—is inherently psychofunctional.

Indeed, Block (1978) seems to make such an identification between information-processing and psychofunctionalism. For Block, whereas analytic functional equivalents to us "need not have … psychological (information-processing) … mechanisms like ours … this … will not apply to … psychofunctional" equivalents (Ibid., 301).

The issue, however, is that there are two different ways of understanding information processing, and only one is proprietarily psychofunctional. We might understand information processing in terms of information to be processed, as Dretske (1981) does. Or we might speak of actual systems processing information, which do so in some sort of language or other medium—a formal system—that represents the information processed. There is a fairly large difference between the two. David Marr (1982) points out that there are multiple formal systems for representing numbers, including Roman and Arabic numerals, and, within the latter category, binary, decimal, hexadecimal, etc. Marr argues that the choice of a formal system has functional upshot: "For example, if one chooses the Arabic numeral representation, it is easy to discover whether a number is a power of

10 but difficult to discover whether it Is a power of 2. If one chooses the binary representation, the situation is reversed. Thus. there is a trade-off; any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover" (Ibid., 21). Thus the informationas-represented in a system is distinct from information-to-be-represented in a system.

The former is proprietarily psychofunctional. What formal system we use to represent information is a matter of empirical fact, and not at all *a priori*. Block's Martian example supports this claim: the difference between the human and Martian is (in part) the language in which they represent premises and conclusions in their inferences, and Block argues that a psychofunctional, but not an analytic functional, theory is sensitive to this difference. But both the human and the Martian are processing the same information, which, per Block, renders them analytic functionally equivalent.

The psychofunctionalist might further objected that the line of thinking I am using to argue for psychofunctionalism's chauvinism can equally demonstrate that analytic functionalism is chauvinistic as well. But abstracting away more than analytic functionalism will encounter the flip side of chauvinism, which Block calls liberalism. To abstract away from analytic functionalism, we must abstract away from state relations and appeal to pure behavioral equivalence. Block (1981) has an example of an organism that is behaviorally equivalent to us, but not minded. He proposes a being with an extremely long list of responses appropriate for a human to give for any given perceptual input. When (e.g.) asked a question, it just queries its list and responds in an appropriate way. Block argues that this being is not minded, and I agree. But it is also not analytic-functionally equivalent to us! For despite being the ‘heir’ of logical behaviorism, analytic functionalism does not posit the inner workings of our minds as a black box. Rather, analytic functionalism involves a rich network of inter-state connections inside our head between various beliefs, desires, etc.

Chapter 6 / Hybrid Functionalism

While psychofunctionalism has drawbacks—and analytic functionalism strengths—that are often overlooked, I do not endorse a full-blown analytic functionalism. In this chapter, I will say why, and explain what I do endorse.

Analytic Functionalism and Explanatory Depth

Given the discussion in the previous two sections, the question arises: why should we not just become analytic functionalists? I do not think we should. For while psychofunctionalism is necessarily chauvinistic, analytic functionalism necessary lacks the explanatory power of psychofunctionalism in three senses. First, psychofunctionalism explains how it is the case that mental states have the analytic functional roles they do. How might a belief, alongside a desire, take perception as input and issue behavior as output? Psychofunctionalism can invoke intermediate systems between perception and belief as well as between belief and behavior that explain those transitions. For instance, a multimodal integration cortex and a concept-subsuming system on the input side, and a motor simulation system on the output side. The invocation of these systems is unavailable to analytic functionalism.

Further, psychofunctionalism can directly explain behavioral differences that analytic functionalism cannot. Suppose humans and Martians perform different inferences at different speeds. Inferences where the translation from English to a formal language is simple are performed much more quickly by the Martians; inferences where the translation is complex are performed much more quickly by us. Analytic functionalism cannot explain these facts for the very reason that the human and Martian are analytic-functionally equivalent!

Finally, psychofunctionalism can explain *breakdowns* in analytic functional roles. Consider Capgras Syndrome, where patients believe[46] that their loved ones have been replaced by imposters. These beliefs are extremely resistant to evidence. All analytic functionalism can do is report the fact that they lack one of the inputs characteristic of beliefs, i.e. being formed or revised in light of new evidence. But psychofunctionalism can appeal to mechanisms that explain why this is the case: Davies et al. (2001) appeal to a malfunctioning system whose role is to remove extremely implausible candidate beliefs

[46] It is still debated whether delusions like those in Capgras Syndrome are *bona fide* beliefs. I believe they are (Bortolotti 2010) but if the reader disagrees, they may substitute here some other preferred example of breakdown in belief functioning.

before they are formed. This system, as part of an empirical psychology, cannot be appealed to by an analytic functional account of belief.

Relatedly, consider a famous experiment by Bruener and colleagues (1956). Subjects are asked to sort cards from a deck of playing cards, and given instructions either to make a pile of [spades and clubs] or [not-diamonds and not-hearts]. Since there are only four suits in a deck of playing cards (let us assume the jokers are removed), these two sets of instructions amount to the same thing. Yet subjects performed the (again, identical) task faster when given the first set of instructions.

On an intuitive level, the reason is obvious: the first set of instructions is simpler. For humans (assumedly) represent positives as 'primitives' in the representational medium in which our thought is realized. That is, we add a negation operator to a basic representation to make it a negative representation, as in P vs. ~P. But this is a fact about representational mediums. Some other minded species, let's say Venusians, could have just as easily evolved to represent negatives as primitives and need some sort of 'affirmation operator' to represent positives.[47] What we think of as (P, ~P) would in the minds of this species be (+P, P). If the explanation for why humans sort cards faster given the instructions to sort spades and clubs, it would follow that Venusians would sort cards faster given the instructions to sort ~diamonds and ~hearts.

[47] I would assume such a species would have been in an environment where the evolutionary pressure that made them develop representations was such that the salience of some negative representation was higher than that of a positive, and thus the former needed to be processed more quickly. For instance, a species that survived better in the absence of some abundant environmental factor, rather than the presence of a scarce one.

A human who represents positives primitively, and a Venusian who represents negatives primitively can be (and let's stipulate, are) relaxed equivalents at the analytic-functional level because the analytic-functional level is the level of computational theory. In terms of computational theory, the subjects—human and Venusian both—are either representing "sort out the spades and clubs from the others" or "sort out the ~diamonds and ~hearts from the others." For the computational theory only specifies the information to-be-represented, not the manner in which it is represented. Yet they are behaviorally different with respect to the sorting task—humas sort faster under one set of instructions, Venusians the other. Because they are (relaxed) analytic-functional equivalents, we can't appeal to analytic functionalism to explain this difference. We can, however, appeal to psychofunctionalism. For psychofunctionalism discusses representations at the algorithmic level, which includes details of the representational language or medium. And that is exactly where we are going to find discussion of the different representational mediums deployed in human and Venusian minds.

Putting Analytic Functionalism and Psychofunctionalism Together

The standard view is that analytic-functional and psychofunctional descriptions are generally taken to be unrelated. Block (1978), introducing the distinction, notes that psychofunctional roles can have as inputs and outputs the same states utilized by analytic functionalism—the difference is that the former is not restricted to these states. Thus, for Block, on a psychofunctionalist view, the posits of folk psychology are not 'cordoned off' and placed at a different hierarchical level. Further, Fodor (1975, 52-3) quite explicitly intends for his psychofunctionalist theory to put folk psychological states and empirically discovered psychological systems at the same hierarchical level.

Block and Fodor's understanding of psychofunctionalism, though it ought to carry significant weight—after all, Block introduced and defined the term, and Fodor was the first philosopher to make a theory to which the label properly applied—is not universal. For instance, Aydede and Fulkerson (2019) discuss 'pain' as having components at an analytic-functional level whose function is explained in terms of systems at a psychofunctional level. This sort of conception involves some sort of hierarchical or mereological relation between the two: beliefs might be constituted by the subsystems discovered by cognitive psychologists, but beliefs (or, more properly, a belief forming or storing system) are not *among* the systems discovered. Within experimental psychology, Barrett (2006; 2017) has proposed a theory of emotions that can be understood along these lines.[48]

We can take this hierarchical conception of functional levels and use it to craft a hybrid theory, where mental state types are explicated in terms of *both* analytic-functional and psychofunctional roles. However, if we are to use this hierarchical conception, then we must i) make a decision as to which level one wants to place beliefs, and ii) explain the

[48] Hirschbach and Bechtel (2015) interpret Barrett this way, and Barrett herself (2009) has said that her view is compatible with the mereological 'mechanistic explanation' posited by Bechtel and others (Bechtel and Richardson 1993/2010; Piccinini and Craver 2011). However, Barrett (2012) herself obfuscates the picture by describing emotions as 'socially constructed), and on Bechtel's mechanistic picture, social constructs exist at a level *above* the states posited in folk psychology (Bechtel 2009). I elsewhere (Pober 2018) argue that Barrett is mistaken to consider emotions socially constructed, and articulate a version of her view where it is clear that emotions stand in exactly the sort of hierarchical relation to psychological systems that Hirschbach and Bechtel read Barrett as saying they do.

nonredundant contribution of the other level. i) is required because no state at the lower level plays the whole functional role of a state at the higher level, so the two cannot be truly identical. Thus the ‘hybrid’ functionalist must choose which state *is* the belief; ii) is required because once we choose, say, the psychofunctional level, we must explain what role the analytic-functional level is playing if my view is to be a hybrid one.

Moreover, i) and ii) are required because belief functional-roles at one level are guaranteed not to be genuinely identical to belief functional-roles at the other level. Recall, the (partial) analytic-functional role of belief involves inputs from perception and outputs to behavior, whereas the psychofunctional role involves more detailed mediating systems in between perception of an object and belief on the one hand, and belief and behavior on the other. The analytic functional role will in a sense be ‘bigger’—that is the state to which this role is attributed will include the state to which the belief is attributed *plus* other psychofunctional states and processes. That the analytic-functional state will include the psychofunctional state is guaranteed by the semantic relation between the two (which obtains whether or not one takes a hierarchical view), since the psychofunctional state is identified as the kind of state it is in virtue of its contribution to fulfilling the analytic-functional role.

Hybrid Functionalism

With that in mind, I need one more conceptual tool in my kit to answer i) and ii). Specifically, I want to invoke Block’s (2007) conception of a ‘core’ versus ‘total’ neural realizer, albeit without the neural part.[49] To use Block's example, the fusiform face gyrus is the neural area to which the function of holistic facial recognition has been localized. But if you took the gyrus out of the context of a working brain, it would not recognize faces even if you ran the right electrical currents through it—it would not do much of anything, really. The FFG is thus the 'core' neural basis—and, if described or identified functionally rather than neurally, the core functional realizer—of facial recognition. That is, it only performs this function in conjunction with other systems, and the set of the FFG and those other systems are the total neural basis—and their functional specifications the total functional realizer—of face perception.

The state of which the psychofunctional role of belief is predicated, plus (in my toy example) the multimodal integration cortex and motor and planning systems are the total realizer of the state of which the analytic-functional role of belief is predicated. Since the multimodal integration cortex and motor/planning systems figure into functions other than belief/desire-caused behavior, they are not the core realizer. What is left—the psychofunctional role-predicated state—therefore *is* the core realizer of the analytic-functional role-predicated state.

Returning to the questions at hand. Regarding i), my view is that the psychofunctional state *is* the belief. But, regarding ii), it is the belief *in virtue of* being the core realizer of the analytic-functional state. And this relation holds between the (singular type of)

[49] Shoemaker (1981) makes a distinction between core and total functional realizers, but my understanding is that this distinction goes beyond what I am looking for. Specifically, the 'total realizer' of a mental state for Shoemaker includes the state of the entire mind of the agent at the time the state is occurrent.

analytic functional-state and every variety of state with a psychofunctional role that can be called ‘belief.’

Beliefs are thus multiply realizable phenomena. This claim is not new (Putnam 1967; Fodor 1974). But I am making it in a different way. Most functionalists posit one ‘functional’ level (though see Lycan 1989) and take states at that level to be multiply realizable with respect to physical states. I am saying that the multiple realizability relation exists *between functional levels* (I will discuss the relation they each have to physical states in the next section).

The way the view maintains explanatory depth is straightforward. It contains all of the depth of a ‘pure’ psychofunctionalism.

Avoiding chauvinism is a bit more complicated, but my hybrid view does so as well. To avoid chauvinism, a view must render an extensionally adequate verdict: it must imply that all minded beings are in fact minded. My view does this just in case all (say) believers have some states or sets of states play the analytic-functional roles of believing.

SECTION THREE: SOLVING PROBLEMS

In the previous section, I articulated a novel functionalism. With its resources at hand, plus our definitions of 'physical' and 'reduction' from section one, we can turn to the ultimate task of this work: to articulate and defend a genuinely reductive theory of mental phenomena. I am especially interested in two problems. First, in Chapter 7, I will examine how to address Block's generality problem. And then in Chapter 8, I will address Multiple Realization, and how to get true nomic equivalence of which I argued Kim's view fell short.

Chapter 7 / Generality and Homeostatic Property Clusters

I have claimed that no extant reductive theory of the mental meets what I've been calling the generality problem. As introduced by Block (1978), this problem amounts to the idea that all members of a mental kind must have something in common. I have also promised that I am going to offer a solution to the generality problem. I therefore need to say what solving this problem amounts to. And it's not quite what one might think *prima facie*.

Nailing Down the Question

Recall that Lewis (1966) distinguished between the sense and reference of a mental concept. According to his (self-labelled) identity theory, the referents of mental kind terms were brain states or properties, whereas the senses were functional role properties.

Lewis couched this distinction in his own descriptive theory of meaning, but we can generalize beyond that. And generalizing beyond that is valuable, since not all theories of meaning agree with Lewis on what constitutes the sense of a predicate or concept. Per Devitt and Sterelny (1999), sense of a natural kind term in a purely causal theory of meaning (i.e., Kripke 1972/1980; Putnam 1975b) is whatever members of the kind share, to be determined by empirical investigation. Yet Lewis is clearly not talking about some underlying 'essence' (in the deflationary sense) but something knowable *a priori*.

To more broadly account for Lewis's distinction between the 'what is it?' and 'in virtue of what is it that sort of thing?' questions, I introduced the idea of a constitutive basis and distinguished between the reference and constitution questions. A typical generalized reduction runs these two questions together. What is water? It is H_2O—all of it, that is, is H_2O. In virtue of what is a mass of stuff water? In virtue of its being H_2O.

Yet the two question need not have the same answer, and this is the key to answering the generality problem. We can give a reductive account of what mental phenomena are. But, as functionalists, we must give a functional account of what makes those mental phenomena (members of) the mental kinds that they are. I will therefore in this chapter be looking to give a functionalist answer to the generality question.

Lewis himself already anticipates this: 'pains' are the disjunction of all physical kinds occupying the 'pain' role (relative to the appropriate species—if a Lewisian Martian somehow has the human neural basis of pain, it would not be pain [Lewis 1980]). But what makes them all 'pain' are their occupying a common functional role.

Answering the generality challenge is giving a univocal answer to the constitution question that refers to something more substantive than a mere predicate. Lewis fails by his own lights because—as I will discuss later in this chapter—functional role properties are insufficiently substantive in the salient sense (Lewis 1983).

This is not to say that the reference question has no conceptual work to do: it is where the charge of multiple realizability needs to be headed off. By ‘multiple realizability charge’ I mean a burden unique to reductive theories: when a reductive theory claims that mental kind M reduces to physical kind P, the mental kind, M, is fixed by an answer to the reference question. For multiple realizability is an issue of what a mental kind is (which I’m calling ‘reference’ here): it is a question of how many physical kinds make up the mental kind in question. Since nonreductive functionalists embrace multiple realizability, they can (as the token identity theorists do) embrace a referent set of multiple physical kinds. (They still have to answer the generality challenge: if all members of a mental kind don’t share what the functionalists claim they do—a functional role—then functionalism is in trouble). Lewis, by identifying the reference set of a mental kind with a disjunct, does not meet the multiple realizability challenge.

Kim separates the questions even more than Lewis, and consequently comes closer to meeting the multiple realizability charge. He splits up general mental kind terms like pain into species- (really, system-)specific sub-kinds that correlate with distinct physical kinds. He thus relieves himself of the burden to answer ‘what is it?’ for pain *as such* and instead only needs to answer it for each system-specific sub-kind of pain. But the move has a cost: he now has to say what all of these distinct physical kinds have in common.

His answer—by his own lights—is only that they share a ‘concept’ or predicate rather than a substantive property that can make them, in his terminology, ‘nomic kinds.’ Alternatively, he argues that his view “may be characterized as the view that psychological kinds have no real essences, only nominal essences” (Kim 2008, 111).

To recap: one can, as Kim does, answer the reference and constitution questions for different mental kinds. The constitution question requires an answer that applies to broad mental kinds (e.g. pain): pain *qua* pain, as, if a view is to meet the generality challenge, it must show that all pain states are pain states in virtue of some common property. The reference question, which requires an answer in terms of a single physical kind if the view is to count as reductive, can be answered for sub-kinds as Kim does. The problem with Kim’s (and Lewis’s) view aren’t that they cheat by splitting the questions apart (and in Kim’s case, answering them for different mental kinds): it is that their answers are insufficient. I will discuss the multiple realizability issue in chapter 8. Here, I want to focus on the constitution question: what suffices for a good answer, and why Kim and Lewis’s answer doesn’t make the cut.

Properties, Sparse and Abundant

For physicalists, both reductive and nonreductive, causation is at the heart of both psychology and all of the sciences ‘below’ it such as biology, chemistry, and physics. We care about the ‘workings’ of the mind because we care about why we think the way we do, and why we act the way we do based on how we think. That is, we seek causal explanations: per Fodor, “what you need in order to do science is a taxonomy apparatus that distinguishes between things insofar as they have different causal properties, and that groups things together insofar as they have the same causal properties" (Fodor 1987, 34).

Structure and function both matter insofar as they speak to causation: (on my picture) structure as underlying it, and function as describing it. Certainly, psychology cares about other aspects of mental life, such as the qualitative and intentional, but the hope of every physicalist is that these can be, if not reduced, then somehow exhaustively explained in terms of the causal powers. It is not for nothing that Jackson (1982) posited qualia as epiphenomenal in order to demonstrate that they resist analysis via physical (spatial/structural/compositional) and functional properties. As Fodor put it: "either … psychology is in the business of causal explanation, or it is out of work" (Ibid, 157n3).

Kim speaks to this concern by focusing on nomic projectible properties (Kim 1998, 108), which are properties that (exemplify kinds which) maintain the same causal powers across all contexts in all nomologically possible worlds. This type of property is also, as Kim (Ibid, 105) notes, the sort of property that exists under a 'sparse' as opposed to 'abundant' conception of properties. I will call properties that exist under a sparse conception 'sparse properties' and all others 'abundant properties' (though this phrasing is a bit awkward). The distinction between sparse and abundant (conceptions of) properties goes back to Lewis (1983), and is worth examining in some detail, as it is shared by Kim as well as other theorists relevant to the current discussion (e.g. Piccinini 2020).

Before doing so, it is worth mentioning another dimension along which conceptions of properties can differ. A categoricalist (e.g. Lewis 1986) takes qualities or attributes (including structural and compositional qualities) to prior to causal powers: an object with a set of qualities will have its causal powers in virtue of those qualities plus nomic laws. A dispositionalist, or advocate of the causal theory of properties (Shoemaker 2001) will see the powers as conceptually prior; the subset view (Ibid; Wilson 2011) is usually explicated in terms of a causal theory of properties (though see Piccinini 2020 as well as my discussion of his view in the previous chapter). Here, I will put my cards on the table and confess to be assuming and working within a categorical view. The aim of science, especially when invoking mechanistic explanation, is to see how physical entities and their properties give rise to the causal powers that they do (in worlds nomologically like ours). Alcohol intoxicates, and it does so because of the chemical structure of ethanol and how that structure interacts with our bodies and brains. A view in which the power of intoxication is held fixed and the chemical structure of the intoxicant altered is technically coherent, but it seems implausible if one is trying to do a philosophical project that is in tune with science. For science, as I see it, is about looking at the basic stuff of the universe—matter (and energy)—and seeing how it interacts to give rise to the world we live in. A causal powers-first view would turn that on its head: matter wouldn't be the basic stuff of the universe!

We should further distinguish between causal powers themselves—endowed in objects by their qualities—and descriptions of those causal powers. Recall the notion of higher-order properties: these include functional roles as discussed previously (Antony and

Levine 1997; Kim 1998; see Morris 2018 for discussion of the role of higher-order properties in other views) as well as dispositions: the disposition of being fragile is the property of having some structural property or properties that endow the property-bearer with the tendency to break when struck with sufficient force (Prior, Pargetter, and Jackson 1982).[50] These properties are descriptions of causal powers rather than powers, or the sort of thing that endows/underlies powers in their own right. This is in part because of overdetermination worries like those described by Kim with respect to functional role properties. Kim's causal exclusion argument doesn't effect functional role properties *qua* functional role properties: it effects all higher-order properties (McLaughlin 2006). As Lewis puts it: "[w]e would not wish to say that the breaking of a struck glass is caused both by its fragility and by the frozen-in stresses that are the basis thereof; and if forced to choose, we should choose the latter" (Lewis 1983, 370n29).

Returning to the main topic at hand: Kim agrees with Fodor's claim that psychology and naturalistic philosophy of mind centers on causal explanations. It is for this reason that he takes up the sparse conception of properties. Per Lewis, the set of sparse properties is those "that there must be to ground the objective resemblances and the causal powers of things, and there is no reason to believe in any more" (Ibid, 345).[51] Whereas abundant

[50] The Prior, Pargetter, and Jackson view is only one view of dispositions, though it is the one I endorse and I believe (independently) the most sensible in the current context. However, Armstrong (1968) takes dispositions to be type-identical with their causal bases. Armstrong's view would render dispositions first-order properties with causal powers in their own right. However, if you do that, then there's no disagreement between Lewis/Kim on the one hand and Shoemaker/Wilson on the other. And the latter are at least intending to disagree with the former.

[51] Lewis is actually in this quote talking about universals, which he then replaces with what he calls 'natural properties.' He does not actually articulate the notion of a sparse or abundant conception of properties, but

properties can correspond to any predicate, no matter how artificial and *post-hoc*. The property of being my dog, a leaf within 30 meters of my apartment, or clothes in my brother's wardrobe is a property according to the abundant conception. These properties are "too disjunctive and[/or] extrinsic to figure in the conditions of occurrence of … the events that cause things" (Ibid 370/370n29).

And this is why Kim, with his concern for projectible nomic properties, endorses the sparse conception (more accurately: he thinks that only sparse properties matter for delineating kinds. He does not necessarily deny the *existence* of abundant properties, as long as we are clear with respect to which properties are and are not sparse). With a nod (footnote) to Fodor, he writes of the sparse conception that "current debates over the mind-body problem and mental causation tacitly presuppose a particularly robust version of this approach according to which differences in properties must reflect differences in causal powers" (Kim 1998, 105).

We can now understand Kim's view that broad (not species- or structure-specific) mental kinds are not real properties. They are descriptions of causal powers, and the sparse properties are to be found in the physical realizers. Since broad kinds do not share a type of physical realizer—we have to go down to species- or structure-specific kinds to find those—they don't correspond to any sparse property. They are thus predicates without properties (see Heil 2003; Morris 2018 for discussion of this relation). Yet they are not

his notion of natural properties is the one that has become what Kim would call properties under the sparse conception.

predicates that express gerrymandered, unrelated disjuncts like some abundant properties. Per Kim, we should

> eschew [sic] the talk of functional *properties* in favor of functional *concepts* and *expressions*. What lends unity to the talk of [higher-order properties like] dormitivity and such is conceptual unity, not the unity of some underlying property. *Q*ua property, dormitivity is heterogeneous and disjunctive, and it lacks the kind of causal homogeneity and projectibility that we demand from kinds and properties useful in formulating laws and explanations. But dormitivity may well serve important conceptual and epistemic needs, by grouping properties that share features of interest to us in a given context of inquiry (Ibid, 110).

Thus, per Kim, concepts have some sort of higher status than just any old predicate, but they do not rise to the level of predicates corresponding with sparse properties. This is Kim's answer to the generality problem. It nonetheless strikes me as insufficient. First of all, 'conceptual unity' has a second-class status at best when compared to causal unity. Although we can intuitively tell the difference between 'pains' and 'my office in Riverside, a particular speck of dust on the moon, and the photons of light illuminating the Eiffel tower at night' there is no ontologically principled difference. Rather, the difference is relative to our 'conceptual and epistemic needs.' And for the generality problem to be met, there needs to be some ontological similarity across all pains.

Science, as Fodor said (and Kim agreed) is in the business of giving causal explanations. In particular, it is in the business of providing generalizations grounded in causal explanations. This link to explanations shows a way in which epistemic concerns might be given some sort of ontological import. What we need is a metaphysics that can capture a tripartite yet principled distinction between sparse properties, properties that are not sparse but still subsume generalizations, and abundant properties.

On further examination, Lewis's distinction between sparse and abundant properties can come closer to doing the trick than my description of it has led on. A quick terminological note: Lewis himself did not predicate 'sparse' or 'abundant' of properties as such. Rather, he posited natural and nonnatural properties, and pointed out that the former are sparse and the latter abundant (Lewis 1983, 345). Lewis's 'naturalness' is directly equivalent to Kim's projectible nomicity: it is about grounding regularities in virtue of homogeneous intrinsic qualities underlying causal powers. What often gets glossed over in more recent treatments of the sparse/abundant property issue is that naturalness, in the relevant sense, comes in degrees. For Lewis, a property is "*perfectly* natural if its members are all and only those things that share some one" intrinsic property that grounds causal powers (Ibid 347). "But we should also have other less-than-perfectly natural properties, made so by families of suitable related [natural properties]. Thus we might have an imperfectly natural property of being metallic" (Ibid). Being metallic seems like the sort of property a science should respect, as it can ground a great many causal generalizations.

The idea of an imperfectly natural property might allow Kim (or an enthusiast of his view) to 'upgrade' functional kinds from mere concepts to (kinds exemplified by) genuine sparse properties. If we count imperfectly natural properties within the realm of the sparse, and we ground the kind in related-but-distinct realizers that endow causal powers, rather than the functional roles themselves, we might be able to work pains as such into the realm of the sparse.

In favor of this suggestion is the fact that nomic projectible properties clearly outstrip Lewis's perfectly natural properties. If 'metal' (the kind exemplified by the property 'being metallic') is not a nomic projectible kind, then I am not sure I have a grasp on the concept. That would invalidate so much of what we consider chemistry that, to paraphrase Fodor from another context, it might well be the end of the world.

But there are considerations that militate against this proposal. First, it's not clear that the diverse realizers of pain across all pain-havers are sufficiently related. More to the point, it's not clear what constitutes being sufficiently related! Nor is it clear what degree of imperfection we should use as the demarcation between sparse and nonsparse properties. Thus the notion of an imperfectly natural property, while barking up the right tree, is too vague to ground the idea that functional kinds are defined by exemplifying real/sparse properties.

The underlying issue, I think, is that while there is a relation between natural (perfect and imperfect) properties and causation, as well as with nomically projectible kinds, they are deployed for very different reasons. Natural properties "should comprise a minimal basis for characterizing the world completely … [those] that do not contribute at all to this end are unwelcome" (Lewis 1983, 346). But Kim's interest in nomically projectible properties is grounded in the motivation he shares with Fodor: to find the properties that underwrite generalizations based on causal regularities that are the bread and butter of science.

The Homeostatic Property Cluster View

Fortunately, there is another project in the ballpark that is quite specifically designed to capture causal regularities: the Homeostatic Property Cluster theory of natural kinds (Boyd 1991; Griffiths 1997). HPC kinds aren't your grandfather's natural kinds—and they certainly aren't Aristotle or Aquinas's. They deny underlying essences—one prominent paper explicating HPC kinds contains "When Traditional Essentialism Fails" in the title (Wilson, Barker, and Brigandt 2007). Instead of grounding kinds in such essences, they ground them in causal regularities. They seek property clusters that are underwritten by causal reinforcement: either an underlying mechanism that causes the properties in the cluster to instantiate, or mutually reinforcing causal influences between the properties themselves (Lipski 2020). The causal relation (of either variety) between properties in a cluster is what makes the cluster count as homeostatic. What is crucial about HPC kinds is that they are aimed at finding the ontological basis of nomic projectibility: they "provide the ontological element of a solution to the problem of induction" (Griffiths 1997, 174).

A debate within advocates of HPC kinds is whether to re-incorporate a watered-down 'mechanistic essentialism'[52]. On this view, HPC kinds not only require an underlying mechanism, they require that members of the kind have *the same* underlying mechanism.

[52] This term comes from personal communication with Matt Barker.

In the past (Pober 2013) I have endorsed such a conception; I would argue that Griffiths (1997) and Craver (2009) as well. Following Craver (Ibid) I will call this version the Mechanistic Property Cluster (MPC) view of natural kinds. Others (Wilson, Barker, and Brigandt 2007; Lipski 2020) allow for multiplicity of mechanisms as long as they do the relevant causal work; Lipski (Ibid) has explicitly argued for this conception. His argument is that since scientists often use functional kinds.[53] Taken at face value, this argument seems insufficient and question-begging for my purposes, since whether functional kinds can count as natural kinds is exactly what I am interested in determining. But taken another way—that functional kinds support the kind of generalizations based on causal regularities that interest scientists—Lipski's argument is more persuasive. For the ultimate arbiter of whether a type of category should count as a natural kind is whether or not it does the work of natural kinds, and Lipski is pointing out that functional kinds often do.

HPC kinds can thus give us a tripartite distinction between MPC kinds (HPC kinds which share a common underlying mechanism), HPC kinds *simpliciter*, and categories that are not natural kinds. Although I am persuaded that HPC kinds should not be restricted to MPC kinds like I previously argued, I still believe MPC kinds have a pride of place. Natural kinds whose property cluster is underwritten by a *common* mechanism are more causally similar than those that are not, and, from an epistemic standpoint, more likely to be the source of causally grounded generalizations yet to be discovered. But they are not

[53] Lipski also provides an argument based on exegesis of Boyd, that this broader conception was what Boyd intended, but I do not consider that here: I am interested in the best view, not necessarily the details of Boyd's.

exhaustive of HPC kinds. To be clear: by ‘mechanistic’ kinds I mean those underwritten by a mechanism as detailed by mechanistic explanation. That is, the mechanism must have parts that are both functionally delineable and spatially/structurally/isolable. Thus, species, the paradigmatic example of natural kinds on the HPC view for Wilson, Barker, and Brigandt as well as Griffiths, aren’t underwritten by a common mechanism if that mechanism is, as Griffiths would have it, a “causal homeostatic mechanism [is] whatever it is that explains the projectability of that category. A microstructural essence is only one kind of … mechanism. Other possibilities include external forces like those produced by the ecological niche of a species … [or] the shared history that holds together the members of a biological taxon” (Griffiths 1997, 212). I disagree: if we are going to relax natural kinds to include multiply realizable/disjunctive mechanisms within a kind, then there is no need to force these weaker/varied ‘mechanisms’ the box of *bona fide* causal homeostatic mechanisms.[54]

Purely functional kinds, like pain, are at least *candidates* for HPC kinds. For any token state or entity instantiating *some* causal basis for a functional role, we know several properties that it instantiates: the various aspects of the functional role, plus any causal roles that necessarily accompany those specified in the functional role (recall the discussion of Ramsification and causal versus functional roles from the chapter 5). And no one is denying that functional kinds have their functional roles in virtue of some

[54] Of course, if, as Devitt (2008) would have it, species are underwritten by DNA sequences unique to each species, *that* would count as a unitary *bona fide* mechanism.

causal/mechanistic basis: the issue is that it is not (necessarily) the same causal basis across realizers.

So we can establish that functional kinds are candidates for being HPC kinds. Are they in fact HPC kinds? This turns, I think, on how many causal regularities are captured in the functional role, and whether those regularities are homeostatic (I am being careful to avoid saying that instances have causal powers in virtue of their functional roles, again, the idea is that their kind membership is based on having a property cluster *described by* the functional role in virtue of some underlying causal mechanism), or, more specifically, how many regularities are required to be captured for a category to count as an HPC kind. For the sort of analytic functional role that defines person-level/folk psychological mental states like belief or pain is rather shallow. For belief, it is largely about their role in practical and theoretical inference: inputs from perception, combining with other beliefs to output new beliefs (in theoretical inference) and combining with desires (and possibly other beliefs) to output intentions (in practical inference). There may be some more features to it, but most of the causal profile of, say, human beliefs, is not necessarily going to be part of the functional role of beliefs that obtains across all species of believers (see chapter 6).

Whether this is enough depends on how many generalizations must be captured by a category to count as an HPC kind. Philosophers writing on the topic—again, I must admit, including myself (Pober 2013) but also, notably Griffiths (1997; Scarantino and Griffiths 2011)—have generally said ‘a good many.’ But I think the key point is that HPC kinds are more than nominal kinds: more can be inferred from an entity’s membership in the kind than the mere fact that it has a single property that makes it part of that kind. As Griffiths (1997) helpfully illustrates: ‘superlunary objects’ is a nominal kind: nothing more can be inferred from an object’s being superlunary than that it exists farther from the Earth than the (Earth’s) moon.

I don’t believe that functional kinds are in danger of being nominal kinds. Even if all the category of beliefs share is ‘the belief role’—which is doubtful given that the functional role is almost always a subset of the causal role of a type of state—they still aren’t mere nominal kinds simply because the specification itself invokes multiple functional roles. If the specification were more simple or atomic—say, doxastic-content bearers rather than beliefs—the category might risk being a mere nominal kind, but we already know that *beliefs* do specific things with the contents they bear.

I thus suggest that the way to solve the generality problem is to replace the sparse conception of properties with the HPC conception of natural kinds. Not only does it allow ‘pains’ to be a natural kind, the shift is not *post hoc*: the HPC kinds project is *designed for the purpose of finding nomic, projectible kinds* whereas Lewis’s natural/nonnatural properties project was not.

It is true that the way I am vindicating the status of ‘pains’ or ‘beliefs’ does not allow for reduction *qua* those categories. But I think Kim is right that “local reductions reduction enough” (Kim 1998, 94). That is, *under certain conditions* (that I will specify in the next chapter) we can have a reductive theory of pains that reduces subtypes of pain while answering the generality by showing that all of these reducible subtypes belong to the same overall type.

Chapter 8 / Multiple Realization, and Why Psychophysical Reduction is (Almost Certainly) True

I don’t think it’s really very plausible to deny that folk psychological mental state kinds have, or at least can have, multiple kinds of physical realizer across nomologically (let alone metaphysically) possible worlds. It is thus incumbent upon the reductionist to say how, given the multiple realizability of such mental kinds, reduction is possible. And reduction has to be a type-type relationship implying nomic equivalence between the two types. In this chapter, I will lay out my response, which takes elements of extant replies by Bechtel and Mundale (1999), Shapiro and Polger (Shapiro 2000; Polger 2004; 2007; Shapiro and Polger 2012; Polger and Shapiro 2016) and Piccinini (2020).

Among nonreductive views that embrace multiple realizability, the subset or aspect realization view, pioneered by Yablo (1992) and taken up and developed by Wilson (2011) and Shoemaker (2001) deserves special consideration for the way in which it powerfully responds to Kim’s causal exclusion argument. I will address this issue here as well, largely basing my reply on remarks made by Morris (2018) and Piccinini (2020).

First, however, I will briefly remark on how I understand the realization relation. It is a relation between tokens of distinct types, subsets of a type, or a type and the subset of another type such that the relata are coreferential in terms of spatiotemporal properties. It is possible for the realization relation to involve or amount to token-identity (*pace* Polger and Shapiro 2016) but it is not required. What is important is that i) there is no spatiotemporal difference in the referents: if the relata in question are a state, they are spatiotemporally coextensive on the token level, if it is properties, they are attributed to spatiotemporally coextensive entity or event tokens, and ii) predicated of the same object. This second criterion excludes Gillett's (2002) notion of 'dimensioned' realization wherein properties of components or parts (e.g. carbon molecules) realize properties of the whole (e.g. the hardness of a diamond). Following Polger (2004; 2007; Polger and Shapiro 2016), I think it is a mistake to call this relation 'realization', rather, it is some type of compositional relation.

Multiple Realizability versus Variable Realizability

Shapiro (2000) suggests a series of scenarios, which I will paraphrase here. In the first one, you have two corkscrews: one painted red and the other painted yellow, which are otherwise identical. Are they distinct realizations of the kind 'corkscrew' (or the property of 'being a corkscrew')? My sense is that no one would say yes unless they are extremely motivated to make the case that the tiniest inconsequential difference results in multiple realizability. On the other hand, take one of these corkscrews (let's say the red one) and pair it with a vacuum suction device. The latter isn't even really a cork*screw*, but both it and the corkscrew might be classified as corked-bottle openers. To the extent that 'corked-bottle openers' is a valid kind (which I will speak to shortly), these two devices are clearly multiple realizations of it.

Other cases fall between these two extremes and are worth discussing in a bit more detail. For the third case, note that there are different kinds of corkscrews. A "waiter's corkscrew" a single lever, usually about the size of a (non-corked) bottle opener that easily fits in a pocket. The same lever is held perpendicular to the screw and rotated around it to turn the screw into the cork. Whereas a "winged" corkscrew is larger and has two levers at a one-hundred eighty degree angle. These levers are attached by a standalone casing for the corkscrew, which has its own top with a handle that can be turned to insert the screw into a cork.

These two corkscrews are multiple realizations of 'corkscrew.' While they are more alike than either one is with respect to the vacuum-based cork remover—they share some parts, such as a screw and a lever/levers to pull the cork out once the screw is in it—they also have many differences.

The fourth case is probably the trickiest, and where I think most peoples' intuitions would run counter to Shapiro's and my own (though I know of no experimental studies probing peoples' intuitions on the topic). Nonetheless, I hope to convince readers that my and Shapiro's interpretation is the correct one. This case is more like the first case, in that you have two corkscrews of the same kind and same measurements (e.g. the length and diameter of the screw—both its cylindrical and spiral diameters) but made of different materials: one is metal and the other is plastic, albeit a fairly strong plastic. Shapiro argues—and I agree—that this is *not* an instance of multiple realization.

What separates the third case from the fourth? Per Shapiro, the differences *make a difference* to how the corkscrew performs its function. The winged corkscrew provides ease of function in the moment—inserting the screw into the cork is less work, and less force is needed on each lever, since the force required to pull the cork out can be split between the two of them—in exchange for being a bigger and more unwieldy instrument (while it's true that someone could make a winged corkscrew with smaller levers, doing so would defeat the purpose, i.e., remove the advantage this type of corkscrew has over the waiter's type). Whereas the metal and plastic waiters' corkscrews material composition does *not* make a difference to the way they function, as long as the plastic and metal both meet some threshold of rigidity. If the plastic were of a weaker sort, such as the kind cheap Happy Meal toys are made of, then the fact that one of the two was plastic would make a difference: the plastic one wouldn't be able to do its job for all but the loosest corks. But if the plastic is strong enough, there is no functional difference between the plastic and metal corkscrews.

The key idea, per Shapiro (see also Polger 2004; Piccinini 2020) is whether physical differences in the realizer make a difference to how the function is realized. There are two related ways of spelling this out that I find helpful. To do so let me introduce the idea of the 'resolution' of a functional description.

Some functional descriptions—notably those posited by analytic functionalism in philosophy of mind—are rather broad and vague, such as a belief combining with other belief via inferential processes (which respect inferential rules) to produce new beliefs. I call this a 'low resolution' function. What sharpening the resolution amounts to, though, will be easier to spell out with the example of a portable radio with the sole function of producing sound as output from radio waves as input. There are two ways in which we can sharpen the resolution. First, we can more precisely describe what the radio does: it receives radio waves only of a certain range of wavelengths, and makes sounds of specific pitches within a certain decibel range. Spelling these out is one way of describing the function at a higher resolution. This aspect of low vs. high resolution is what Bechtel and Mundale (1999) call coarseness (low res) or fineness (high res) of grain. In another sense, when we sharpen the resolution on the radio, we can see that much of the matter composing it has nothing to do with its function: the reception of radio waves is performed entirely by a receiver, perhaps with the help of an antenna (I'm really dating myself here, aren't I?). The sound output is produced entirely by a speaker. And they are connected by some basic electronics which perform a translation function. This aspect of sharpening resolution amounts to a mechanistic decomposition. And of course we can combine these two elements, by making more precise a description of the parts.

Determining whether or not a physical difference in realizers amounts to a difference in performing the function *a la* Shapiro amounts to whether the functional descriptions of the two token devices (/mental states) becomes different at a higher resolution, along either dimension. Shapiro's case for the metal and plastic corkscrews being of the same type is based on the fact that it does not. We can describe the function at a finer grain, we can break down the function of the corkscrews into the functions of their parts, and we will not find a difference (we can find a difference if we iterate the decomposition many steps down, to the interactions between plastic molecules and metal molecules, but this does not make a difference to the functioning of the parts at the N-1 level, which is what matters [Shapiro 2000; Polger 2004; Polger and Shapiro 2016]).

When the physical differences among distinct realizers make a difference to how the functional token is realized, we have genuine multiple realizability. When they do not, we have what Piccinini (2020, 53) calls “variable realizability.” I will finish this chapter by arguing that variable realizability is no threat to reduction, but there is more work to be done before I can make that case.

I suspect some are still not convinced: after all, Shapiro made these arguments two decades ago, and not all of the defenders of multiple realizability are simply ignorant of his work. I will thus offer a diagnosis of the sticking point via an alteration on the example, call it 4B. In this case, we have two metal corkscrews which are identical aside from being made of different metals: one is steel and the other is aluminum. The difference between 4B and ‘4A’ is that steel and aluminum are sub-types of a kind with which we are all familiar—metals—whereas a metal and a plastic are not.

For those who still believe the two metal corkscrews are an instance of multiple realizability, I take it that they associate multiple realizability with distinct ‘stuffs.’ Steel and aluminum are different stuffs, thus two otherwise identical (functionally defined) objects realized in them are multiply realized. But this is an outdated and unscientific conception of natural kinds. For metal is a natural kind. It is a natural kind because it does what natural kinds do: it possesses commonalities in its physical composition that endow similar if not identical causal powers, and these causal powers are similar enough to explain a great many causal generalizations. Natural kinds can exist at multiple levels (Body 1991): thus genuses can be natural kinds as much as species can. There is no reason to reject metals from the circle of natural kinds.

'Metal-and-plastic' as such do not form a natural kind (or, at least, I am not going to defend here the claim that they do). But metal and the type of plastic in our imagined corkscrew from example 4 *do*, I think, form a natural kind. It is not a natural kind we have previously identified, but they have a common compositional property or set of properties that endow the two devices with many identical causal powers. Specifically, their *rigidity* (Shapiro 2000) endows them with these causal powers. Not all plastics are this rigid: indeed, I would think that some of the softest metals like copper or bronze do not themselves possess enough rigidity to be effective corkscrews, especially if the handles of the levers are made out of pure samples. Thus this kind does not involve all metals or plastics. But it involves a subset of both, and perhaps some other materials (petrified wood would likely make a good corkscrew). Solids with X rigidity, while a kind created *post hoc* does the work of a natural kind, and there is thus no scientific reason to fail to treat it as one. What *kind of* kind these somewhat gerrymandered categories make is something I will address shortly, but I will first need to discuss subset/aspect realization.

Subsets and Aspects

Aspect realization is a special subclass of multiple realization. The realized property or state stands in a determinable-determinate relation with respect to its realizers. This means that the properties that define the realized property—such as ‘having four sides’ for a quadrilateral—are an aspect or subset of the properties that define the realizer properties. In addition, the properties of the realized property are properties all of the realizers must share. I am choosing to call this the ‘aspect’ view following Piccinini (2020). He points out that the notion of ‘subset’ has some implications that I do not believe are properly part of this view: “sets are abstract objects, and … to [some] philosophers, appealing to sets commits proponents of the subset view to the existence of sets *qua* abstract objects” (Ibid, 27). The issue here is that if sets are abstract objects, then a subset is not a proper part of a set: it is a distinct being. But there is no reason why we should understand the relation between the essential properties of a quadrilateral and a square that way.

As I noted in chapter 3, the aspect view offers a response to the causal exclusion problem that vindicates the causal efficacy of the mental. But it does so at a cost. The kind of properties that define a square in addition to a quadrilateral—i.e., having equal sides, having right angles—are not *different in kind* from the property that defines a quadrilateral in the first place. They are all geometric properties. Likewise, as Yablo (1992) acknowledges, the properties that define the determinable mental type are, like the properties that define its determinate physical realizers, *physical properties*. If they are physical properties, then aspect realization does not violate reduction as long as the type of reduction in question is mental-to-physical (Morris 2018; Piccinini 2020).

It does falsify psycho*neural* reduction, which is the aim of some nonreductive physicalists (e.g., Antony 2003; see Appendix 1). But I am defending psychophysical, rather than psychoneural, reduction. For I see the former as the only reduction that has a chance of being true: mental kinds can't reduce to neural kinds because they can exist in beings without neurons. But mental kinds *can* reduce to kinds that are made of neurons and whatever realizes mental phenomena in those other beings: kinds based on what neurons and those other realizers have in common. Indeed, a common thread among the views whose authors claim—and I contest—nonreductive status is that they are defining nonreductive as denying psychoneural, rather than psychophysical reduction. I see no reason to privilege psychoneural reduction: after all, I am not just talking about human (or terrestrial) mental phenomena.

Returning to the topic of the kind 'solids of X rigidity' we can now define that kind as exactly the type of kind the aspect view invokes. It is a kind where the various members share an aspect—in this case, the structural property of having X (or >X) rigidity. Thus the kind is not only one that does the job of natural kinds, it is a perfectly mundane sort of kind: an aspect kind.

Multiple Realization among Functions

In the previous section, I have discussed some of the various resources available to the reductionist to diffuse the multiple realizability objection. However, I need to add one more component to this analysis to make a plausible case that psychological kinds are reducible. This is because at least some psychological kinds rather clearly do not meet the criteria spelled out here.

It is extremely implausible that belief, understood as the state playing an analytic functional role characteristic of belief, is not multiply realizable. The criteria are so minimal—playing into practical and theoretical inference, taking input from perception, etc.—that it is highly unlikely that there is only one mechanistic decomposition that holds for all such states across all believers in the whole range of nomologically possible worlds.

Consider the functional kind, ‘word producer,’ that is, some device—biological organ or artifact—that produces words for the range of beings that can recognize words, be they competent minded users of a language or mere talk-to-text programs that we all use and tolerate. I can immediately think of two ways to divide this category that (at least when combined) resist common mechanistic explanation. First, there is a division between visual and audio word producers: devices that produce patterns of markings that we can recognize as words versus devices that produce sound waves that our ears (or a microphone) can record and interpret as words. Can they be said to have common mechanisms, or even any common parts? Perhaps they both have some sort of pattern generator for letters, but the patterns underlying sounds and written letters are so different—both at the level of individual letter and in the sense that sound patterns are less rigid (more work is done by the interpreter to discriminate, e.g., /b/ and /p/, or where one word starts and another ends). But there is another distinction that, when combined with the audio/visual one, obliterates any commonality ‘pattern generators’ might hold. Here I speak of the distinction between analog and digital. Thus, there can be analog-audio, digital-audio (a distinction that applies to LP’s vs. CD’s and audio files), analog-video, and digital-video. When we handwrite letters, we are producing words in analog fashion: the words on a computer screen or projector as well as those made by a dot-matrix printer are digital. But ‘digital’ is equivocal when it refers to audio and visual signals: digital-video is made up of dots, whereas digital-audio is made up of individual ‘bits’ of sound.

Given these basic differences—and that I’m sure there are many more on further investigation—it just seems extremely unlikely that there is a univocal word-producing mechanism. Of course—keeping with the analogy that ‘word-producer’ is a stand-in for analytic functional roles of psychological kinds—it would be great for my view if these extremely abstract roles *were in fact* underwritten by univocal mechanisms. If there is such a mechanism for the analytic functional role of belief, then my work is done! But I do not think there is.

Consider the sub-(sub-) category of digital-video word producer. Here we have computer screens (of the LCD and CRT variety), projection screens, and dot-matrix printers (both laser and inkjet). This is still a pretty diverse crowd! But here I think the prospects for a univocal mechanistic explanation, or something near enough, are much better.

The unifying element is that all digital-video word producers involve a pixelated dot-matrix function. A dot-matrix can be spatially defined: it is basic nodes or pixels in a grid formation that cover a segment of two-dimensional space. The dot matrices across devices are quite different: the dots are ink for a printer, bits of liquid for an LCD monitor, and bits of light projected onto a screen for CRT monitors and standard projectors. The pixels you see and the device that causes them to appear are one and the same in an LCD monitor but not any of the other devices. Of the projected pixels, those made by a printer are (relatively) permanent compared to projectors/CRT devices, where the projected pixels disappear once the light source is turned off. But they all share a significant aspect: a grid of nodes that can be lit up or held dark (or potentially lit up in various colors, but since we're dealing with word, and not picture, projectors, we can ignore that detail). That they share such an aspect means they have a lot in common above and beyond their ability to project 'written' words. They break down in the same way: the part of the device responsible for illuminating the pixels malfunctions leaving little of the visual field blank—analog visual word producers do not break in this way. The sort of 'code' that gives them instructions can be fairly similar, with instructions on which pixels to illuminate and which to leave dark—again, analog video devices do not work this way.

Given this significant commonality, we can ask: are digital video word projectors a multiply realizable kind? I think they are not, but more work needs to be done to explain why. I said that not being multiply realizable involves having a single underlying mechanistic explanation across the kind. Digital video word projectors do not have a single, univocal mechanistic explanation. They all have a common physical aspect—a pixel/dot matrix display—but that is not sufficient to say that the whole underlying mechanism is common across dot matrix printers, LCD screens, etc.

However, they all share a common mechanistic component part, but this common part is embedded in otherwise different mechanisms, be it different (other) component parts, a different organization among parts, or both. In an LCD screen, the display and the 'projector' are one and the same part, whereas in a dot matrix printer, the 'projector' is really an ink propulsion device that puts the ink on paper, and the paper is the display.

Recall that our interest in multiple realizability issued from its ability to prevent reduction of a kind. While the overall mechanistic picture underlying digital video word projectors is multiply realizable, whether the kind is reducible turns on how the kind itself relates to the component that is shared throughout its extension.

The Reduction of the Mental (Finally!)

Let's return to discussing beliefs. Suppose they are like digital video word projectors: a common mechanistic component embedded in otherwise distinct mechanisms (I will examine whether or not beliefs are actually like this shortly). If we understand beliefs as the state which occupies the analytic functional role characteristic of believing, then beliefs are not reducible in virtue of having a common mechanistic component across the whole category.

But I have argued that we *shouldn't* understand beliefs as the whole occupant of the analytic functional role. In articulating Hybrid Functionalism, I have argued that we should understand them as the *core realizers* of those occupants. Core realizers are a mechanistic component, albeit a privileged one. They are the component that is unique to believing, whereas other components of the mechanism underlying the occupant of the whole analytic functional role might also be components underlying the occupant of the analytic functional desire role. Pixel/dot matrix displays are the core component of digital video (word) projectors. And they are a physical kind, despite being embedded among other mechanistic parts that are quite distinct (e.g., a paper spool for the printer, a light source for CRT and projection monitors, etc.). If beliefs are nomically equivalent to single physical kind then beliefs are reducible. Consequently, if beliefs are the core realizer of the occupant of the analytic functional role, and that core realizer is nomically equivalent to a single physical kind, *then beliefs are reducible*.

But the whole set of occupants of the analytic functional role of believing is, I would argue, much more like word projectors than digital video word projectors. I will now argue that it can still be reducible.

The key conceptual tool here is the notion of low or high resolution functions. Recall that there are two distinct senses in which we can 'raise' the resolution of a function: we can describe it more precisely (at a finer grain) or we can go down a mechanistic level. Let's deal with the former case now.

A coarse-grained function is itself multiply realizable in terms of finer-grained functions. In this scenario, we can 'offload' the multiple realizability between functional kind and physical kind to a relation among functional kinds.

This consideration leads to my final definition of reduction:

> Reduction$_{\text{(final def)}}$ = X type-reduces *qua* nomological type to Y iff i) Y is a physical type, ii) X inherits its causal powers from Y, and iii) descriptions of the properties of Y univocally explain descriptions of the properties of X. Or, X type-

> reduces *qua* nomological type if X is multiply realized in *functional* categories P, Q, and R, and P, Q, and R, each meet i)-iii) for some Y.

I can now spell out the promissory note I made in the introduction. Some mental kinds are directly reducible to physical kinds. The range that are so reducible is greatly increased when we take into consideration i) what is required for something to be a physical kind, especially that kinds can be ‘metal’ rather than only ‘aluminum,’ ‘copper,’ etc., and ii) that for many folk psychological kinds, the actual state itself is the core realizer of the analytic functional role. And while the whole occupant of an analytic functional role’s being reducible implies that its core realizer is reducible, the core realizer can be reducible even when embedded in different mechanisms, as in distinct digital video word projectors.

But psychophysical reduction generally can still hold even if not all mental kinds are reducible even with the bag of tools listed above to enhance reduction. For many psychological kinds, *qua* functional kinds, are multiply realizable *in terms of higher resolution functional kinds*. The more we sharpen the resolution, the more we get to a point that any physical difference in the realizer either makes a difference to the realized function—in which case we can keep sharpening the resolution—or it does not, in which case the physical realizers constitute a unified physical kind. And while it is possible for two radically different physical bases to realize the *exact* same physical kind, it is extremely unlikely.

Kim, Redux

Returning to the generalization problem, the set of reducible belief sub-kinds form an overarching functional kind that counts as a *bona fide* natural kind per the HPC view. I think that the local reducible versus overarching irreducible functional kinds is what Kim is getting at when he accepts ‘local’ or species-/system-relative reductions (higher resolution) as opposed to global functional concepts (lower resolution). But his way of getting there doesn’t quite work, for he can’t cash out how the overarching kind earns its causal-explanatory keep given his sparse conception of properties. For Kim, psychology is species-/system-specific, so that for each functional kind within a system, there is a single reduction base. But making psychology specific in this way throws in the towel on the generalization problem and is thus unacceptable for my purposes. Without the system-specificity, however, there is no guarantee that two species will not have the same functional kind in distinct physical realizers. What Kim needs to augment his picture is the aspect view plus the notion of coarser/finer grains of function. With those two tools combined, I can say that *if* two species/systems have identical functional kinds *at a fine grain*, then their physical realizers will at least share a common aspect.

Now we bring in the second way of sharpening the resolution of a function. If the aspect the (more finely grained) functional roles share is their core realizer, then we can—as Hybrid Functionalism says we ought to for independent reasons—identify the mental kind in question with the core realizer, and not the whole occupant, of the functional role. Putting it all together: if two species have different physical realizers of a common functional role, then the kin designated by that functional role is reducible if *either* a) the functional kinds that correspond with beliefs in each species are actually different functional kinds at a finer grain, or b) the species have states with identical finer-grained belief functional roles. In case b), they will share *at least* a core realizer *qua* physical kind (or the whole occupant *qua* physical kind, in which case they also necessarily share the core realizer).

What if two species/systems do have identical belief-functional roles at the highest grain of resolution, yet share no physical aspect of their realizers (core or total) in common? Then the kind is not reducible. This makes the reduction of any given functional kind an empirical matter (albeit a difficult one, as the empirical aspect of the question turns on what can exist in any nomologically possible world!)

Nonetheless, given the number of resources I (or Kim) can deploy, this scenario is unlikely. For any coarse/'medium' grained functional role, the difference in physical realizers must, first of all, go beyond mere variable realizability. Because variably realizable physical kinds make up a single, 'ur'-kind, like metals. If the difference does go beyond mere variable realizability, then it cannot correspond with different functional roles at a finer grain of functional description, like different kinds of word projectors. And if it does go beyond mere variable realizability, and it does so with respect to functional roles at the finest grain of description, then we must rule out identifying the mental/psychological kind with the core realizer of the functional role and not the whole occupant. Yet I have given independent reason to think we should identify mental kinds with the core realizers.

While I thus cannot rule out the irreducibility of beliefs, I can conclude that it is much less likely than their reducibility. If beliefs are like word projectors, I can divide up the kind into finer grained functional roles, like digital video word projectors. If the mechanism isn't the same across the more finely grained functional kind, it likely still has enough similarity so that I can point to an aspect and/or the component that is the core realizer.

There is one further point to be made here. Recall Kim's point that any physical difference implies a causal difference. Most physical differences that seem to constitute multiple realizability of a single functional role really, at a finer grain of description, turn out to be different functional roles. Consider my earlier example of various devices that propel flying transports. A propellor, jet engine, and rocket all meet the bill. But they are all doing very different things. A rocket propels by burning fuel directly. Propellors and jets require fuel to operate it, but they do not propel *by* burning fuel. They require fuel as an energy source, rather than a propellant, and thus could be 'fueled' by a battery rather than gasoline. Jets and propellors both function by creating an air pressure differential that creates forward momentum by making the air behind the engine higher pressure than the air in front of it, but they do so quite differently. Propellors, like the fixed wings of airplanes, use Bernoulli's principle to create a pressure differential with air in front of the propellor being lower because it has to travel a further distance to get 'past' the propellor blade. Jet engines do this by compressing the air inside the engine and then expelling it from the rear.

The upshot is that for two physical kinds to realize the *exact same* function is an extremely high bar. They can't have minor physical differences, as that will result in either the same function in virtue of the physical differences being irrelevant to the function (variable rather than multiple realizability), or it will result in similar but not quite identical functions. Rather, they must come at realizing the function from *radically different* physical starting points.

An objection to the analysis I am giving might be: genuine multiple realizability that is not "offloadable" to a multiple realization relation between functions at various grains is extremely rare. Yet multiple realizability seems ubiquitous. I agree with this point; I do not agree that it is a criticism. I am not even saying I am biting the bullet: if I am biting anything, it is a nice, soft bit of food. The fact is that once reduction is properly defined, and once multiple realizability is properly understood, the latter just isn't a very plausible obstacle to the former. I am not alone in this conclusion. Per Shapiro (2000, 636): "It is my contention here that philosophers are much too quick to claim that a given kind is multiply realizable. Once various conceptual issues are clarified, the task of demonstrating multiple realizability smacks the hard surface of empirical fact, and, I shall argue, leaves MRT far more difficult to establish than philosophers currently acknowledge."

Chapter 9 / Conclusion

Why is the belief that reduction is impossible so widespread? A large part of the culprit is that philosophers think reduction implies that physics captures all interesting empirical generalizations, and thus entails elimination of all explanatory vocabularies above physics. But integrating reduction with mechanistic explanation shows why this is simply not the case. Beyond that, I can only speculate as to the reasons why reduction remains unpopular despite having such a solid foundation. Kim (likewise speculating) makes an intriguing point:

> Expressions like “reduction,” “reductionism,” “reductionist theory,” and “reductionist explanation” have become pejoratives not only in philosophy, on both sides of the Atlantic, but also in the general intellectual culture of today. They have become common epithets thrown at one’s critical targets to tarnish them with intellectual naivete and backwardness. To call someone “a reductionist,” in high-culture press if not in serious philosophy, goes beyond mere criticism or expression of doctrinal disagreement; it is to put a person down, to heap scorn on him and his work (Kim 1998, 89).

I agree with Kim. I think the tendency is especially strong in philosophy with respect to reduction of *the mental* as it goes against the idea that we are doing something that science cannot do. I think this is misguided: philosophy and science are complementary, with neither at risk of being fully replaced by the other. It also goes against the idea that mentality is essentially rational in a normative sense: here Davidson (1970) was right that the two are incompatible if aimed at the same task. But the conclusion I draw from their incompatibility is that mental normativity and the causal powers of the mental are simply two different things: there is how we describe what the mind does, and then there is how we evaluate it, and we should stop trying to force our philosophical evaluations into the way the mind works in its own right.

Bibliography

Adams, F. (1979). “Properties, functionalism, and the identity theory.” *Eidos* 1:153-179.

Antony, L. (2003). “Who’s Afraid of Disjunctive Properties?” *Philosophical Issues* 13:1-21.

Antony, L. (2008). “Multiple Realization: Keeping it Real.” In J. Hohwy and J. Kallestrup (Eds.) *Being Reduced: New Essays on Reduction, Explanation, and Causation*. New York: OUP. 164-175.

Antony, L. (2015). “Reality and reduction: What’s really at stake in the causal exclusion debate.” In T. Horgan, M. Sabates and D. Sosa (Eds.) *Qualia and Mental Causation in a Physical World: Themes from the Philosophy of Jaegwon Kim.* Cambridge, UK: Cambridge University Press. 1-24.

Antony, L. and Levine, J. (1997). “Reduction with Autonomy.” *Philosophical Perspectives* 11:83-105.

Armstrong, D.M. (1968). *A Materialist Theory of Mind*. New York: Routledge.

Armstrong, D.M. (1968/1994) *A Materialist Theory of Mind: Second Edition*. New York: Routledge.

Arpaly, N. and Schroeder, T. (2013) *In Praise of Desire*. OUP

Aydede, M. and Fulkerson, M. (2018). “Reasons and Sensory Affect.” In D. Bain, M. Brady, and J. Corns (Eds) *Philosophy of Pain*. London: Routledge 27-59.

Barrett, L.F. (2006) ""Solving the Emotion Paradox: Categorization and the Experience of Emotion." *Personality and Social Psychology Review* 10:20–46.

Barrett, L. F. (2009). "The Future of Psychology: Connecting Mind to Brain." *Perspectives in Psychological Science* 4:326–39.

Barrett, L.F. (2012). "Emotions are Real." *Emotion*, 12:413.

Barrett, L.F. (2017). "The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization." *Social Cognitive and Affective Neuroscience* 12:1–23.

Bechtel, W. (2007). "Reducing Psychology While Maintaining its Autonomy Through Mechanistic Explanations." In M. Schouten and H. Looren de Jong (Eds.) *The Matter of Mind.* Malden, MA: Blackwell. 173-198.

Bechtel, W. (2009) "Mechanism, modularity, and situated cognition." In P. Robbins and M. Aydede (Eds.) *The Cambridge Handbook of Situated Cognition*. Cambridge, UK: Cambridge University Press, 155-170.

Bechtel, W. and Mundale, J. (1999). "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66:175-207.

Bechtel, W. and Richardson, R.C. (1993/2010) *Discovering Complexity*. Princeton, NJ: Princeton University Press.

Block, N. (1978). "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9:261-325.

Block, N. (1981). “Psychologism and Behaviorism.” *Philosphical Review* 90:5-43.

Block, N. (2007). “Consciousness, Accessibility, and the Mesh Between Psychology and Neuroscience.” *Behavioral and Brain Sciences* 30:481-538

Block, N. and Fodor, J.A. (1972). “What Psychological States are Not.” *The Philosophical Review* 81:159-181.

Boyd, R.C. (1991). “Realism, Anti-foundationalism, and the Enthusiasm for Natural Kinds.” *Philosophical Studies* 61:127–48.

Braddon-Mitchell, D. and Jackson, F. (1999). “The Divide and Conquer Strategy to Analytical Functionalism.” *Philosophical Topics* 26:71-88.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: Wiley.

Burge, T. (1979). “Individualism and the Mental.” *Midwest studies in philosophy* 4:73-121.

Burge, T. (1993). “Mind-body Causation and Explanatory Practice.” In J. Heil and A. Mele (Eds.) *Mental Causation*. New York: OUP. 97-118.

Carruthers, P. (2006). *Architecture of Mind.* New York: OUP.

Chalmers, D. (1990). “Why Fodor and Pylyshyn were wrong: The simplest refutation.” *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* 340-47.

Chalmers, D. (1996). *The Conscious Mind*. New York: OUP.

Chisholm, R. (1966). “Freedom and Action.” In K. Lehrer (Ed.) *Freedom and Determinism*. New York: Random House. 11-40.

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Clarendon/OUP.

Craver, C.F. (2009). “Mechanisms and Natural Kinds.” *Philosophical Psychology* 22:575-594.

Cullison, Andrew. 2010. “What Are Seemings?” Ratio 23:260–74.

Cummins, R. (1975). “Functional Analysis. *The Journal of Philosophy*, 72:741-765.

Davidson, D. (1969). “The individuation of events.” In *Essays in honor of Carl G. Hempel*. Springer, Dordrecht 216-234. Reprinted in Davidson, D. (1980) *Essays on Actions & Events.* New York: Clarendon Press 163-80.

Davidson, D. (1970). “Mental Events.” Reprinted in D. Davidson (2009) *The Essential Davidson* 105-118.

Davidson, D. (1982). “Rational animals.” *dialectica, 36*(4), 317-327

Davies, M., Coltheart, M., Langdon, R., and Breen, N. (2001). “Monothematic Delusions: Toward a Two-Factor Account.” *Philosophy, Psychiatry, and Psychology* 8:133-158.

Dennett, D.C. (1969). *Content and Consciousness*. New York: Routledge.

Dennett, D.C. (1978). “Where Am I?” in Dennett, D.C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press 356-64.

Dennett, D.C. (1991). *Consciousness Explained*. New York: Little Brown.

Dennett, D.C. (1995). “Superficialism vs. Hysterical Realism.” *Philosophical Topics* 22:530-36.

Devitt, M. (2006). *Ignorance of Language*. New York: OUP.

Devitt, M. (2008). “Resurrecting Biological Essentialism.” *Philosophy of Science* 75:344-382.

Devitt, M. and Sterelny, K. (1999). *Language and Reality, 2nd Edition.* Cambridge, MA: MIT Press.

Dretske. F.I. (1981). *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.

Dretske. F.I. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Egan, F. (1992). “Individualism, Computation, and Perceptual Content.” *Mind* 100:461-84.

Feigl. H. (1958). “The ‘mental’ and the ‘physical.’” In H. Feigl, M. Scriven, and G. Maxwell (Eds.) “*Minnesota Studies in the Philosophy of Science* 2:370-497.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Festinger, L., Riecken, H., & Schachter, S. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.

Fodor, J.A. (1974). “Special Sciences: Or, The Disunity of Science as a Working Hypothesis.” *Synthese* 28:97-115

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

Fodor, J.A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Gendler, T.S. (2008). “Alief and Belief.” *Journal of Philosophy* 105-634-63.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

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

Griffiths, P.E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Griffiths, P.E. and Scarantino, A. (2009). “Emotions in the Wild: The Situated Perspective on Emotion.” In P. Robbins and M. Aydede (Eds) *The Cambridge Handbook of Situated Cognition*. Cambridge UK: Cambridge University Press 437-53.

Hacking, I. (2007). “Putnam's theory of natural kinds and their names is not the same as Kripke's.” *Principia: an international journal of epistemology* 11:1-24.

Heil, J. (2003). *From an Ontological Point of View*. New York: OUP.

Hemmo, M. and Shenker, O. (2022). “Why Functionalism is a Form of Token-Dualism.” In S. Ioannidis, G. Vishne, M. Hemmo, and O. Shenker (Eds.) *Levels of Reality in Science and Philosophy: Re-examining the Multi-level Structure of Reality.* Springer Nature Switzerland: Cham, Switzerland. 115-152.

Herschbach, M. and Bechtel, W. (2015). “Mental Mechanisms and Psychological Construction.” In L.F. Barrett and J.A. Russell (Eds) *The Psychological Construction of Emotion*. New York: Guilford Press 21-44.

Jackson, F. (1982). “Epiphenomenal Qualia.” *The Philosophical Quarterly* 32:127-36.

Jackson, F. (1998). *From Metaphysics to Ethics*. New York: OUP.

Jackson, F. (2012). “In defence of the identity theory mark I.” In S. Gozzano and C.S. Hill (Eds) *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge, UK: Cambridge University Press, 150-66.

Jackson, F., Pargetter, R. and Prior, E.W. (1982). “Functionalism and Type-Type Identity Theories.” *Philosophical Studies* 42:209-225.

Jackson, F. and Petit, P. (1993). “Folk belief and commonplace belief.” *Mind and Language*, 8:298-305.

Kammerer, F. (2015). “How a Materialist Can Deny that the United States is Probably Conscious—Response to Schwitzgebel.” *Philosophia* 43:1047-57.

Kim, J. (1973). “Causation, Nomic Subsumption, and the Concept of Event.” *The Journal of Philosophy* 70:217-236.

Kim, J. (1989). “Mechanism, Purpose, and Explanatory Exclusion.” *Philosophical Perspectives* 3:77-108.

Kim, J. (1992). “Multiple Realization and the Metaphysics of Reduction.” *Philosophy and Phenomenological Research* 52:1-26.

Kim, J. (1993). “Mental Causation in a Physical World.” *Philosophical Issues* 3:157-176.

Kim, J. (1995). “Mental Causation: What? Me Worry?” *Philosophical Issues* 6:123-151.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kim, J. (2008). “Reduction and Reductive Explanation: Is One Possible Without the Other?” In J. Hohwy and J. Kallestrup (Eds.) *Being Reduced: New Essays on Reduction, Explanation, and Causation*. New York: OUP. 93-114.

Kripke, S. (1972/1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Levin, J. (2018) “Functionalism” entry in *Stanford Encyclopedia of Philosophy*.

Levy, N. (2016). “Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements.” *Nous* 49:800-23.

Lewis, D.K. (1966). “An Argument for the Identity Theory.” *Journal of Philosophy* 63:17-25.

Lewis, D.K. (1970). “How to Define Theoretical Terms.” *Journal of Philosophy* 67:427-46.

Lewis, D.K. (1972). “Psychophysical and Theoretical Identifications.” *Australasian Journal of Philosophy* 50:249-258.

Lewis, D.K. (1980). “Mad Pain and Martian Pain.” in N. Block (Ed.) *Readings in the Philosophy of Psychology, Vol. I*. Harvard University Press, 1980, pp. 216-222.

Lewis, D.K. (1983). “New Work for a Theory of Universals.” *Australasian Journal of Philosophy* 61:343-387.

Lewis, D.K. (1986). *On the Plurality of Worlds*. Malden, MA: Blackwell.

Lipski, J. (2020). “Natural diversity: A neo-essentialist misconstrual of homeostatic property cluster theory in natural kind debates.” *Studies in History and Philosophy of Science Part A, 82*, 94-103.

Locke, J. (1824). *An Essay Concerning Human Understanding* 12th Edition. London: Rivington.

Loewer, B. (2002). “Comments on Jaegwon Kim’s *Mind in a Physical World*.” *Philosophy and Phenomenological Research* 65:655-662.

Lycan, W.G. (1979). “A New Lilliputian Argument Against Machine Functionalism.” *Philosophical Studies* 35:279-87.

Lycan, W.G. (1981). “Form, Function, and Feel.” *The Journal of Philosophy* 78:24.

Lycan, W.G. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Machamer, P. Darden, L. and Craver, C.F. (2000). “Thinking about mechanisms.” *Philosophy of Science* 67:1-25.

Madva, A. (2016). “Why Implicit Attitudes are (Probably) not Beliefs” *Synthese* 193:2659-84.

Mandelbaum, E. (2014). *The Architecture of Belief: An Essay on the Unbearable Automaticity of Believing*. University of North Carolina, Chapel Hill, Dissertation.

Mandelbaum, E. (2016). “Attitude, Inference, Association: on the Propositional Structure of Implicit Bias.” *Nous* 50:629-58.

Mandelbaum, E. (2019). “Trouble with Bayesianism: An Introduction to the Psychological Immune System.” *Mind & Language* 34:141-57.

Marr, David. 1982. *Vision*. New York: W.H. Freeman.

McDowell, J. (1994). *Mind and World*. New York: OUP.

McLaughlin, B. (2006). “Is Role Functionalism Committed to Epiphenomenalism?” *Journal of Consciousness Studies* 13:39-66.

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

Morris, K. (2018). *Physicalism Deconstructed*. Cambridge: Cambridge University Press.

Melnyk, A. (2003). *A Physicalist Manifesto*. Cambridge, UK: Cambridge University Press.

Nagel, E. (1961). *The Structure of Science*. New York: Harcourt, Brace & World.

Oppenheim, P. and Putnam, H. (1958). “The Unity of Science as a Working Hypothesis.” In H. Feigl, M. Scriven, and G. Maxwell (Eds.) “*Minnesota Studies in the Philosophy of Science* 2:3-36.

Papineau, D. (2008). “Must a Physicalist be a Microphysicalist?” In J. Hohwy and J. Kallestrup (Eds.) *Being Reduced: New Essays on Reduction, Explanation, and Causation*. New York: OUP. 126-148.

Pargetter, E.W., Prior, R., and Jackson, F. (1982). “Three Theses About Dispositions.” *American Philosophical Quarterly* 19:251-257.

Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition.* New York: OUP.

Piccinini, G. (2022). “Physicalism: Flat and Egalitarian.” In S. Ioannidis, G. Vishne, M. Hemmo, and O. Shenker (Eds.) *Levels of Reality in Science and Philosophy: Re-examining the Multi-level Structure of Reality.* Springer Nature Switzerland: Cham, Switzerland. 195-208.

Piccinini, G. (ms) "Neurocognitive Mechanisms: Some Clarifications."

Piccinini, G. and Craver, C.F. (2011). "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183:283-311

Pober, J.M. (2013) "Addiction is Not a Natural Kind." *Frontiers in Psychiatry* 4:123.

Pober, J.M. (2018). "What Emotions Really Are (In the Theory of Constructed Emotions)" *Philosophy of Science* 85:640-59.

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

Polger, T.W. and Shapiro, L.A. (2016). *The multiple realization book*. New York: OUP.

Prior, E. W., Pargetter, R., & Jackson, F. (1982). "Three theses about dispositions." *American Philosophical Quarterly,* 19(3):251-257

Putnam, H. (1963). "Brains and Behavior." In R.J. Butler (Ed) *Analytical Philosophy: Second Series*. Malden, MA: Blackwell 24-36.

Putnam, H. (1967). "The Nature of Mental States." Reprinted in H. Putnam (1975) *Mind, Language, and Reality*, Cambridge: Cambridge University Press 429-440

Putnam, H. (1975a). "Philosophy and our mental life." In H. Putnam (1975) *Mind, Language, and Reality: Philosophical Papers Volume 2*. Cambridge, UK: Cambridge University Press.

Putnam, H. (1975b). “The Meaning of ‘Meaning.’” *Minnesota Studies in the Philosophy of Science* 7:131-93.

Pylyshyn, Z.W. (2003). *Seeing and Visualizing: It’s Not What You Think*. Cambridge, MA: MIT Press.

Quilty-Dunn, J. (2020). “Perceptual Pluralism.” *Nous* 54:807-38.

Quilty-Dunn, J. and Mandelbaum, E. (2018). “Against Dispositionalism: Belief in Cognitive Science.” *Philosophical Studies* 175:2353-72.

Rescorla, M. (2020). “The Computational Theory of Mind” entry in *Stanford Encyclopedia of Philosophy*.

Rey, G. (1997). *Contemporary Philosophy of Mind*. Malden, MA: Blackwell.

Rey, G. (2003). “Chomsky, Intentionality, and a CRTT.” In L. Antony and N. Hornstein (Eds.) *Chomsky and His Critics*. Malden, MA: Blackwell 105-39.

Ritchie, J. B. (2019). “The content of Marr’s information processing framework.” *Philosophical Psychology, 32*(7), 1078-109.

Ryle, G. (1949). *The Concept of Mind*. New York: Barnes & Noble.

Saidel, E. (2009). “Attributing mental representations to animals.” In R.W. Lurz (Ed.) *The philosophy of animal minds*. Cambridge, UK: Cambridge University Press 35-51.

Scarantino, A. (2014). “The Motivational Theory of Emotions.” In D. Jacobson and J. D’Arms (Ed.s) *Moral Psychology and Human Agency*. New York: OUP 156-85.

Scarantino, A. and Griffiths, P.E. (2011). “Don’t Give Up on Basic Emotions.” *Emotion Review* 3:444-54.

Schouten, M. and Looren de Jong, H. (2007). “Mind Matters: the Roots of Reductionism.” In M. Schouten and H. Looren de Jong (Eds.) *The Matter of Mind.* Malden, MA: Blackwell. 173-198. 1-28.

Schroeder, T. (2004). *Three Faces of Desire.* OUP

Schwitzgebel, E. (2002). “A Phenomenal, Dispositional Theory of Belief.” *Nous* 36:249-75.

Schwitzgebel, E. (2012). “Mad Belief.” *Neuroethics* 5:13-17.

Schwitzgebel, E. (2013). “A Dispositional Approach to the Attitudes: Thinking Outside the Belief Box” in (Ed.) N. Nottlman *New Essays on Belief*. Palgrave.

Schwitzgebel, E. (2015). “If Materialism is True, the United States is Probably Conscious.” *Philosophical Studies* 172:1697-1721.

Schwitzgebel, E. (2016). “Is the United States Phenomenally Conscious? Reply to Kammerer.” *Philosophia* 44:877-83.

Schwitzgebel, E. (2019) “Kant Meets Cyberpunk.” *Disputatio* 55:411-435.

Schwitzgebel, E. (forthcoming). “The Pragmatic Metaphysics of Belief.” In D. Kindermann and C. Borgoni (Eds.) *The Fragmented Mind*. New York: OUP.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science* 77:477-500.

Shapiro, L.A. (2000). “Multiple Realizations.” *The Journal of Philosophy* 97:635-654.

Shapiro, L.A. (2010). “Lessons from Causal Exclusion.” *Philosophy and Phenomenological Research* 81:594-604.

Shapiro, L.A. (2017). “Mechanism or Bust? Explanation in Psychology.” *British Journal for the Philosophy of Science* 68:1037-1059.

Shapiro, L.A. (2022). “Rethinking the Unity of Science Hypothesis: Levels, Mechanisms, and Realization.” In S. Ioannidis, G. Vishne, M. Hemmo, and O. Shenker (Eds.) *Levels of Reality in Science and Philosophy: Re-examining the Multi-level Structure of Reality*. Springer Nature Switzerland: Cham, Switzerland. 209-228.

Shapiro. L.A., and Polger, T. (2012). “Identity, variability, and multiple realization in the special sciences.” In S. Gozzano and C. Hill (Eds.) *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge, UK: Cambridge University Press. 264-287.

Shoemaker, S. (1981). “Some Varieties of Functionalism.” *Philosophical Topics* 12:93-119.

Shoemaker, S. (2001). “Realization and Mental Causation.” *The proceedings of the 20th world congress in philosophy* 9:23-33.

Siegelmann, H.T. and Sontag, E.D. (1995). “On the Computational Power of Neural Nets.” *Journal of Computer and System Sciences* 50:132-50.

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

Soames, S. (2002). *Beyond rigidity: The unfinished semantic agenda of naming and necessity*. New York: OUP.

Sternberg, S. (1969). “Memory Scanning: Mental Processes Revealed by Reaction-Time Experiments.” *American Scientist* 57:421-457.

Strawson, G. (1994). *Mental Reality*. Cambridge, MA: MIT Press.

Tumulty, M. (2011). “Delusions and dispositionalism about belief.” *Mind & language*, *26*(5), 596-628.

Wilson, J. (2006). “On Characterizing the Physical.” *Philosophical Studies* 131-61-99.

Wilson, J. (2011). “Non-reductive Realization and the Powers-based Subset Strategy.” *The Monist* 94:121-54.

Wilson, R.A, Barker, M.J., and Brigandt, I. (2007). “When Traditional Essentialism Fails: Biological Natural Kinds.” *Philosophical Topics* 35:189–215.

Woodward, J. (2005). *Making things happen: A theory of causal explanation.* New York: OUP.

Woodward, J. (2008). “Mental Causation and Neural Mechanisms.” In J. Hohwy and J. Kallestrup (Eds.) *Being Reduced: New Essays on Reduction, Explanation, and Causation*. New York: OUP. 218-262.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, 101:245-280.

APPENDICES

I have removed some text from the main body of the dissertation that is relevant to the discussion, and arguably enhances it, but is not essential for my arguments to go through. I reproduce these excerpted sections here.

Appendix 1 / Antony and the Limits of Nonreductive Physicalism

Antony's (1999; 2003; 2015; Antony and Levine 1997) is the first of two views I will discuss that claim not to be reductive but ought to be classified as such. Of the two, hers is the more ambiguous—the case I will make for Piccinini is fairly straightforward, and one he essentially agrees with (Piccinini 2022). Nonetheless, Antony's view is sophisticated and, in many ways, compelling: it is thus worthy of the same status—of being discussed in detail—as the other four views explicated in this chapter.

Recall that when Kim opted for local functional reductions, which posited nomic equivalence between subsets of a functional kind and their univocal physical realizers, he did so by choosing it over the idea that a whole functional kind was nomically equivalent to its disjunct of physical kinds of realizers. Per Kim, a disjunct of physical kind "fails to represent a uniform set of causal powers, and it seems to drop out of the picture in favor of its realizers. I think we might as well be straightforward and identify [instances of subsets of a functional kind with their specific realizers], bypassing unwieldy disjunctions” (Kim 2008, 109).

Here we find Antony’s departure from Kim. She argues that we ought to identify mental-cum-functional kinds with the disjunct of their physical realizer kinds, on the basis that these disjuncts *can* have uniform causal powers, i.e., be projectible. Per Antony:

> In German, if you wish to speak of a single member of the family of domesticated bovinae, without committing oneself to the beast’s gender, you can, I am told, use the single word ‘kuh.’ But to do it in English, you must circumlocute: since ‘cow’ refers only to female bovids, and ‘‘bull’’ to male, you must say something like ‘member of the family of domesticated bovinae’, or if your Latin’s rusty, ‘cow or bull’ (Antony 2003, 9-10).

The moral of the story is that “the mere fact that English speakers have only a (complex) disjunctive term to do what German speakers can do with a primitive term has no bearing on the relation between the properties the two terms express” (Ibid, 10). And this moral—that whether we happen to have a primitive predicate to refer to a putative property is not the ultimate arbiter of whether this putative property is *bona fide*—generalizes. Antony then puts her cards on the table: “What I really think is that there is no such thing as a ‘disjunctive property’—rather, there are only disjunctive predicates. Moreover, I think that properties are, in themselves, neither ‘mental’ nor ‘physical’—there are only mentalistic and physicalistic predicates, which may or may not express real properties” (Ibid, 9).

Her claim is that mental predicates pick out real properties. Real properties, for Antony, are ones that underwrite projectible predicates: “here’s a short answer to the question of when disjunctive predicates express nomic properties: when they are necessarily co-extensive with projectible predicates" (Ibid, 12). Antony's point is that a disjunctive predicate in one kind of scientific vocabulary (set of proprietary predicates, as Fodor would have it) may be a primitive, projectible predicate in another. Thus 'alcohol' is a projectible predicate in (organic) chemistry because it refers to various compounds with salient structural similarities that can be projected onto any new chemical we are told is an instance of alcohol, even though they are disjunctive at a lower-chemical (perhaps inorganic chemistry?) level between methanol, ethanol, isopropyl alcohol, and so on. Antony's claim is that mentalistic predicates designate real kinds (or properties the exemplification of which constitutes a kind): in short, they are like alcohol.

But then the question of how it is that mental kinds constitute real properties/kinds arises. There are two ways Antony can go here. And, while she seems ambivalent between them, the difference between the two has major implications for both the plausibility and the nonreductive status of her view.

Antony understands projectability in epistemic terms: "Psychological predicates, like macrophysical predicates, geometric predicates, biological predicates, and economic predicates are used by us to state generalizations, predications and explanations that afford us real epistemic power" (Ibid). It is the supporting (essentially epistemic) generalizations, predictions and explanations that make a property projectible. In a similar vein, she claims "[t]he projectability of … predicates entails, at a minimum, that certain epistemic agents (us, typically) have epistemic access to things grouped in particular ways. So, the projectability of mentalistic predicates entails that we human beings have epistemic access to certain phenomena grouped together as mental" (Antony 2015, 23).

But defining projectability of predicates in epistemic terms glosses over a crucial difference between two ways of capturing generalization-subsuming phenomena. On the one hand, we can *describe* a regularity, on the other hand, we can refer to some underlying cause of the regularity. Crucially, only the latter way guarantees that there is such a cause: sometimes patterns appear randomly. More importantly, mere descriptions do not refer to properties that exist on the sparse conception of properties, even if they correspond with such properties. Take a common mentalistic functional role (really an aspect of one), the tendency of perceptual experiences to lead to the entokening of perceptual beliefs. Surely there is some causal power that perceptual experiences (or some system(s) intermediate between them and beliefs) have that underwrites this entokening. That causal power is a property that exists on the sparse conception, the tendency itself—the functional role—does not.

Which does Antony think the nomic property referred to by mentalistic predicates is—the descriptive functional role or the causally efficacious underlying property? At times she seems to go both ways. On the one hand, she says "the generalizations of folk psychology are, epistemically speaking, realization independent" (op cit.), which suggests the former reading. If this is right, then her view is nonreductive, since it insists that functional roles are distinct, either in extension or in causal powers, from their realizers. But then at best her disagreement with Kim is really about whether properties that do not exist on the sparse conception are scientifically ‘real’ or nomic. They certainly fail to be natural kinds, for the reasons discussed in the previous paragraph.

On the other hand, she often talks as if she is fine with the identification of functional roles with *some sort* of physical property that can confer real causal powers. At times she sounds like a subset theorist: “The causal powers of M are not identical with the causal powers of any particular realization base; rather they are identical with the intersection of the causal powers of all the realizers” (Antony 2003, 18). Further, she is willing to accept the claim that mentalistic predicates refer to physical properties! Per Antony: “[s]uppose, for the moment, that we bite the bullet and simply … accede [sic] to the weak reductionist thesis that mental properties are physical properties. But so what? We can still maintain that mental properties are distinct from any of the disjunct properties—we can deny, for example, that mental properties are identical with neurological properties—and that’s the reductionist thesis that Kim is pushing” (Ibid, 7) and that philosophers of mind/psychology “want nothing more for our proprietary properties than the defenders of biology and geology want for theirs” (Antony 2015, 9).

This is clearly reductionist on my understanding of the term. It may be ‘weak’ reductionist, but—to perhaps stretch a metaphor a bit—you can mock an athlete in the weight training room as ‘weak,’ but you have to shut your mouth if they get it done on the field. And what Antony is calling weak reduction gets it done on the field, or so I argue. Further, if she wants the ‘autonomy’ of psychology to be that of biology, then she just wants it to count as yet another mechanistic level, albeit one that might be realized in different mechanisms across various kinds of creatures. But mechanistic explanation guarantees some sort of structure-function correspondence, whereas full-on nonreductive physicalism, as token identity, requires no pattern of relations between the two.

Appendix 2 / Schwitzgebel vs. Lewis on Mad Mental States

One immediate benefit of hybrid functionalism is that it allows a more attractive solution to the question of whether a 'mad' state—one without any aspect of its characteristic analytic functional role—is possible than either Lewis or Schwitzgebel offer.

Schwitzgebel (2013) gives the following example of a putative believer. Suppose that, according to some (otherwise quite plausible) theory of belief, Andi believed that giraffes are six feet tall. Yet:

> Andi is not at all inclined to act and react in the usual way. She is not at all disposed … to say that baby giraffes are six feet tall. If asked explicitly, she would say giraffes are probably born no more than three feet tall. If shown a picture of a giraffe as tall as an ordinary man she would assume it's not a newborn. … And so forth, robustly, across a wide range of actual and counterfactual circumstances. None of these facts about Andi are due to the presence of *ceteris paribus* defeaters like guns to her head or manipulation by evil neuroscientists or a bizarre network of other attitudes like thinking that "three" means six … [But i]f Andi's dispositional structure is like *that*, she doesn't believe that giraffes are born six feet tall [no matter what the theory says]" (Schwitzgebel 2013, 83-4).

Schwitzgebel couches his argument in disposition-talk. But this is easily translated into talk of functional roles:[55] what Andi is disposed to do in virtue of her belief is due to the causal powers of that belief, which, on my way of understanding functions, are tantamount to its functional role (or at least the output end of it).

The example of Andi is intended to be a *reductio ad absurdum* of the notion of a belief that *entirely* lacks its characteristic analytic functional role. And indeed, Schwitzgebel argues that beliefs must maintain some aspect of their functional role on the token level, a view he calls 'token functionalism' (Schwitzgebel 2012).

Lewis disagrees with Schwitzgebel, and regularly states that the generalization need not apply to all tokens of a type. But he only once articulates a view of how Andi's mental state might count as a belief that giraffes are six feet tall, in his (1980) "Mad Pain and Martian Pain" (that Schwitzgebel's account is intended to directly contrast with Lewis's can be gleaned from the title of the paper where he names his 'token functionalism': "Mad Belief?"). According to

Lewis (or, at least, extrapolating his account of pain to belief), Andi's mental state can count as a belief that giraffes are six feet tall despite its lacking any aspect of the functional role characteristic of such a belief *iff* it is of the same physical state type that has that role in typical members of a population (e.g. species) of which Andi is a

[55] Schwitzgebel (2002) considers dispositionalism and functionalism to be at least closely related views. He explicitly compares his dispositional view of belief to Armstrong's theory, which he calls "the classic treatment of mental states as dispositionally specifiable" (Schwitzgebel 2002, 272n11).

member.[56] Lewis couches his view in his identity theory, but that commitment can be easily relaxed for the nonreductive physicalist: just replace the mental state's *being* of the same physical type with *being realized by* a physical state of the same physical type.

I think Lewis is right about the general principle: it seems like we should allow for exceptions at the token level. Yet I do not think Lewis's account of when we should allow those exceptions works—I think Schwitzgebel is quite right about Andi, even if they have the right kind of underlying physical state—but I will hold off on explaining how until I have introduced my hybrid functionalism.

What I want to emphasize for the moment is what Lewis's account does get right: that if we are going to say why a token state lacking an analytic functional role is a token of a type defined by that role, then we need some other way of specifying the state in question to be of the same type of state as that which typically possesses the relevant role. This is because specifying the state in other terms is required for us to be able to tell a story about every exception to the rule. And *that*, in turn, is what is required if a rule with exceptions is to count as a rule that is necessarily true.

The problem with Schwitzgebel's token functionalism lies in its exceptionlessness. While Schwitzgebel comes up with an example—Andi's putative belief about the height of giraffes— that would be absurd to call a belief with that content, it seems premature to rule out *all* such cases. But Lewis's solution won't work either. Per Lewis, the mental

[56] Lewis does not require the population to be a species: rather it has to be some sort of natural kind of population, *like* a species. I ignore this detail for the purposes of the current discussion.

state must be of (or realized by) the same physical state type as that which typically has (or realizes) that role in typical members of the species Andi belongs to. What this gets wrong—and what Schwitzgebel gets right—is that what makes some physical state (the realization basis of) a mental state is what *it* does at the token level, not some similarity relation it bears to other physical states. After all, functionalism (Lewis's reductive version or otherwise) at its heart is about defining mental states in terms of what they do.

Earlier, I noted that Lewis was right that *any* account on which a state defined by an analytic functional role lacked that role on a token basis of this sort would need some alternate method of specifying states. There must be some way to say the token state in question is *like*—of the same type—as states which have that analytic functional role without appealing to the role itself. Seen in this light, the problem with Lewis's solution is that specifying the state physically won't do.

Hybrid functionalism gives us multiple ways of specifying a state functionally. A state with a given psychofunctional role can fail to be the core realizer of an analytic functional role based on malfunctions in other (non-core) parts of the realizer. I suggest that there can be 'mad' states of some mental state type S iff there potentially exist states that have (the appropriate species' version of) S's psychofunctional role which fail to realize any aspect of the analytic functional role characteristic of S.

For a psychofunctional state to fail to realize any aspect of the analytic functional role it does in a properly functioning mind, it would seem that everything else has to go wrong.

Psychofunctional states often have intermediates between them and the ‘surface’: for instance, in between a desire and behavior lies motor planning and action selection systems. These might be part of the ‘total realizer’ of the analytic role of desire. If a desire’s psychofunctional role is intact, then its connection to the very next system ‘in line’ must be intact—otherwise it lacks the psychofunctional role. Suppose that the ‘next system in line’ after the (psychofunctionally specified) desire state is an action selection system. That connection needs to be intact. But for that state to fail to realize the analytic functional role, some *other* connection between action selection and bodily movement would have to be malfunctioning. A ‘mad’ state would mean such a malfunction applies to *every* aspect of a characteristic analytic functional role, otherwise it would realize *some aspect* of the analytic-functional role.

What makes Schwitzgebel’s example of Andi compelling is that it illustrates quite vividly how much is involved (inferentially, behaviorally) in the analytic functional role characteristic of any belief. For Andi to have the belief on my account, they would need to have an ‘intact’ psychofunctional state consequently, how implausible it is that so much is going wrong in Andi’s mind *outside of* the psychofunctional belief itself. Perhaps, then, Schwitzgebel is right that mad *belief* is impossible (though I do not want to commit to this claim). But the ‘everything else is going wrong’ scenario might be less implausible for a state with a more ‘compact’ analytic functional role—perhaps (to use Lewis’s own example), pain.

Appendix 3 / Defending Analytic Functionalism

Part One: Platitudes and Paralytics

Ned Block, in his influential paper "Troubles with Functionalism" (1978) makes two powerful criticisms that apply to all versions of analytic functionalism—Lewis's, Shoemaker's, and mine alike. Explicating and responding to these criticisms will occupy the remainder of this chapter. In both cases, I argue that Block's criticism turns on attributing fewer resources to the analytic functionalist than are truly available to her. Determining what these resources are, and emphasizing them will allow me to finish crafting the strongest version of analytic functionalism.

Block's first critique has to do with the over-emphasis on behavior he sees in analytic functionalism. If an individual lacks the ability to behave in certain ways, she thereby lacks mental states that have those behaviors in their respective functional roles. His main example is that of a paralytic: for such a person, any generalization having to do with a behavior output (as opposed to a state transition output) will be false. Take the (very likely true) generalization that pain, combined with a series of other mental states (such as the desire to avoid pain) leads to a behavior involving avoidance of that pain. Block says that a paralytic can have all of the mental states without typically producing the behavioral output: as he puts, it "But in the case of a … paralytic, attempts to locomote … might *typically fail"* (Block 1978, 297). Indeed, it seems that such attempts would *always* fail to produce the behavioral output (unless her motility is somehow later restored).

Block takes analytic functionalism to imply that the paralytic will be unable to have (feel) pain. If true, this implication would certainly be a problem for analytic functionalism. But I am not so sure that it is. The analytic functionalist seems to me to have two possible responses.

The Scope of Typicality

First, the analytic functionalist could argue that Block is mis-reading the scope of 'typically.' When Putnam (1963) criticizes Ryle's (1949) analytic behaviorism, he uses the example of "Super-Spartans." Super-Spartans are beings with brains and bodies like ours who have somehow trained themselves to show no behaviors (including utterances) typically associated with pain. For Ryle, the criticism is devastating. Ryle renders mental phenomena to be logical constructs of behaviors. The behaviors can be counterfactual: if I am in great pain, I may not scream in pain if I'm in the company of someone I'd like to impress, but the fact that I *would* have screamed had this person not been there is enough for Ryle to rightly say I'm in pain.

The Super-Spartans, though, would not scream (or do anything else) in *any* circumstances. Yet Super-Spartans have pain by hypothesis. Putnam accounts for their pain by defining pain (and other mental phenomena) as the typical *cause* of these behaviors, not, as Ryle (1949) would have it, a logical construct of them. Here, Putnam is using 'typical' to apply to that which generally happens in most organisms of a kind; 'typicality' predicates both intraorganism and interorganism patterns. The Super-Spartans, per Putnam, are sufficiently like us in terms of their brain structure that we can say they have the brain state that typically causes pain behaviors in organisms *like* them, even if not in them.

But then why can we not say the same of paralytics, simply substituting functional states for brain states? They are in pain when they have the (functional) state that plays the pain-behavior-causing role in most organisms like them. Block seems to be assuming that for a platitude to be typically true, it must (typically) happen in *all* organisms of the same kind (i.e. a species).

There is a limit to how broadly this move can be deployed, though. It can't simply be deployed *ad hoc*; there must be a story as to why the putative pain state in the organism in question isn't fulfilling the functional role it typically does in organisms of that type. This burden can, too, be met. Tumulty (2011) argues there are two ways of making such a claim non-*ad hoc*. It requires either i) an excuse, or a story about why the link between cause and effect was 'masked' during a particular instance (but would stand under most other counterfactual circumstances) or ii) an explanation, or a story about why the link fails to instantiate in the organism in question in the first place. The difference, as I understand it, is the same as the two scopes of 'typicality' that I discussed earlier: excuses are for the specific instances where an effect fails to manifest in an organism where it generally does, and explanations are for the instances where the organism, not the event, is the exception. My not yelling out in pain when breaking a bone in order to, say, impress a date would count as an excuse (though I am not sure why this would be impressive); a Super-Spartan's never yelling out in pain because she had trained herself not to would be an explanation. The platitude between being in pain and yelling out in pain would be preserved assuming every instance where someone did not have pain just in case each token event where pain was present but the pained person did not yell out involved either an excuse or an explanation in Tumulty's sense.

Further, this response is limited in that it can only be deployed by the analytic functionalist who has a secondary method of picking out mental states. That is, she must have some method of saying how a given state that is not causing pain behaviors in the paralytic is *the same kind of state* as the state that is causing pain behaviors in the rest of us. Saying 'it is the state that typically causes pain behaviors' is no help. Since it is not causing pain behaviors in the paralytic, that description will not suffice to pick it out of her psychological stew. My next response will suggest a way that an analytic functionalist can do just this, but the fact that I need to appeal to a distinct argument to make this one work demonstrates its limits.

The 'Mentally Paralyzed'

Block's example emphasizes the behavioral outputs of various mental states. But mental states also have outputs that are other mental states. For example, a belief that P, in the presence of a belief that Q (or vice-versa) may have an 'output' of a belief that P&Q. All of these outputs, which I'll call 'cognitive outputs,' are going to be preserved in the paralytic. It would seem to be an empirical question whether *most* of the outputs of a given mental state type are going to be behavioral or cognitive. But it would also seem to be a conceptual point that all outputs for few, if any, states will be either exclusively behavioral or cognitive. Most mental state types will have a healthy mix of both as paradigmatic outputs.

For this reason, the paralytic isn't an example of someone lacking a *vast* majority of the outputs typical of a given state type. In other words, as long as there are 'cognitive outputs' which are preserved in the paralytic, then the analytic functionalist can i) say why the state in question is 'pain,' and ii) maintain that the pain state retains a large chunk of its constitutive functional role.

But what if someone really were lacking the vast majority or entirety of any attitude's typical outputs? Consider a person who was mentally as well as physically 'paralyzed.' In addition to not being able to act on her beliefs in combination with her desires, she would *also* be unable to make inferences from her beliefs. Not only could she not go to the fridge and get a beer if she wanted one based on her belief that there is beer in the fridge, but she couldn't infer that she didn't need to tell her roommate to take beer off the shopping list. She couldn't even infer that there was beer somewhere in her house!

I don't think tweaking the example helps Block here. For if *this* is the sort of case we are talking about, then I would say that the person *really does not have the belief!* Schwitzgebel (2013) lays the case for this claim out with an elegant example that I discussed in the previous appendix. Suppose that, according to some (otherwise quite plausible) theory of belief, Andi believed that giraffes are six feet tall. Yet:

Andi is not at all inclined to act and react in the usual way. She is not at all disposed, for example, to say that baby giraffes are six feet tall. If asked explicitly, she would say giraffes are probably born no more than three feet tall. If shown a picture of a giraffe as tall as an ordinary man she would assume it's not a newborn. If a zookeeper were to tell Andi that giraffes are born six feet tall, Andi would feel surprised and would say, "Really? I would have thought they were born much smaller than that!" And so forth, robustly, across a wide range of actual and counterfactual circumstances. None of these facts about Andi are due to the presence of *ceteris paribus* defeaters like guns to her head or manipulation by evil neuroscientists or a bizarre network of other attitudes like thinking that "three" means six … [But i]f Andi's dispositional structure is like *that*, she doesn't believe that giraffes are born six feet tall [no matter what the theory says]" (Schwitzgebel 2013, 83-4).

Anchoring Functionalism

Block can concede that someone without the ability to make inferences lacks beliefs, but nonetheless maintain there is a plausible case to assume paralytics *simpliciter* are a problem for analytic functionalism. The idea is that *something* more than mere internal state relations is required for a state to count as a belief: there must be some sort of relation to the outside world. Block (1978) calls this the problem of 'anchoring' functionalism. If simply having the right internal relations were sufficient for mentality, then all sorts of random entities, like a country's economy that just happened to realize the part of the Lewis-Ramsey sentence dealing with internal relations would be considered minded (this point is an extension of Block's general charge that analytic functionalism relaxes sufficient conditions for mentality too much--a point I discuss in detail shortly.).

However, the anchoring Block discusses has an input and an output side. A paralytic is still 'anchored' to the outside world via the perceptual input relations she has with it. Is being half anchored sufficient? Strawson (1994) thinks so. He gives the example of 'weather-watchers' a species of immobile beings who closely watch the weather. They are, per Strawson, quite happy when it is sunny and quite sad when it is rainy (potentially, there are even contrarian weather watchers who would prefer Seattle to Los Angeles). In order for them to be this way, the weather watchers must by definition have some sort of perceptual system that allows them to discriminate between sunny and rainy weather and emotions that are elicited in response to the weather. We can add a little more complexity to get them processes that would require the attribution of beliefs and desires to explain. Specifically, they would need the ability to think about weather in the future. If they could think about the weather tomorrow, given that they would be happy if it were sunny, they would then desire it to be sunny tomorrow.[57] And if they can notice patterns, i.e., cloudy today means rain is more likely tomorrow, then they would have predictions--a species of belief—about what the weather would be like tomorrow.

The example of the weather watchers—or perhaps, better, the example of the weather-predicting watchers—and the intuitive plausibility of their having beliefs and desires, demonstrates that being anchored to the external world exclusively by input is sufficient.

[57] Strawson, as well as Tim Schroeder (2004; Arpaly and Schroeder 2013) assume the weather watchers *automatically* have desires, without needing any capacities other than those described by Strawson. I disagree, and think they need some sort of future-weather representational apparatus. For the purposes of the current discussion, though, nothing substantive turns out the point: Schroeder, Strawson, and I all agree that weather watchers plausibly have desires and beliefs.

And the same would likely be true of beings who have only limited sensory capacity—say, only vision, and only black-and-white vision at that.

What if a being had neither input nor output anchoring, but still could make inferences between internal states? Would that being have attitudes? Here, I think our intuitions simply fail us. On the one hand, someone who was born with sensory (and possibly motor) abilities who then lost them wouldn't intuitively lose their internal states upon losing their anchors. We can plausibly picture a sense-deprived paraplegic still performing addition, contemplating philosophy that she had read before being deprived of her sense abilities, and so on. Dennett (1978) vividly illustrates an example of a Brain-in-a-vat who temporarily ends up in this condition--on return to being anchored (in this case with motor as well as sensory abilities), he describes the experience as extremely unpleasant. He (the example is supposedly Dennett himself) describes it as like being nowhere, but not as ceasing to think or exist.

What if someone were born without either sensory or motor abilities? Here I would be inclined to at least *prima facie* doubt their possession of beliefs and desires. But I do not think that any of our intuitions really speak to this example. We simply cannot imagine what it would be like. Further, I do not think this particular case helps Block. The upshot of this discussion is: all that is needed for having attitudes is most of the internal inter-state relations described by the Lewis-Ramsey sentence of analytic functionalism, and just a bit of anchoring (i.e. a few input-output relations described by the Lewis-Ramsey sentence with perception as the input). And that upshot is enough to say that someone who never instantiates a good many of the relevant input-output relations is not a counterexample to defining mental states/mentality generally in terms of the Lewis-Ramsey sentence of analytic functionalism.

The Upshots

There are two upshots to take away from this section. The first, which issues from my first response to Block, is simply that we need to emphasize that the generalizations of analytic functionalism are not and need not be true of all minded beings at all times.

The second response to Block's criticism, however, is the more important for furthering our understanding of analytic functionalism. The basis of Block's criticism is that analytic functionalism is untenable because it ties functional roles too closely to behavior. Indeed, Block says: regarding "specification of inputs and outputs, [analytic] Functionalists require externally observable classifications (e.g., inputs characterized in terms of objects present in the vicinity of the organism, outputs in terms of movements of body parts). Psychofunctionalists, on the other hand, have the option to specify inputs and outputs in terms of internal parameters" (Block 1978, 269). Consequently, only "Psychofunctionalism has the option of using empirical-theory construction in specifying inputs and outputs so as to draw the line between the inside and outside of the organism in a theoretically principled way" (Ibid, 273).

Block is surely right in that the whole Lewis-Ramsey sentence of analytic functionalism must give commonsense (pretheoretic) characterizations of some inputs and outputs characterized this way. But surely there are also mental states/variables within the Lewis-Ramsey sentence who are defined entirely in terms of relations to other states as well.

Why would anyone think otherwise? I think the error—and it is a common one—is to place too much emphasis on where analytic functionalism is (metaphorically) located in the dialectic of (20th century, Anglophone) metaphysics of mind. As Block rightly notes, analytic functionalists "are the heirs of logical behaviorists" (Ibid., 268) like Ryle (1949). For those theories, as I discussed earlier, it was true that every mental state needed to be defined directly in connection with some 'external' phenomenon. But analytic functionalists, while acknowledging this lineage, explicitly emphasize that one of the ways their view improves upon its predecessor is by allowing functional connections to other mental states.

Among analytic functionalists, Armstrong comes the closest in defining functional roles as necessarily having behavioral outputs. For he claims that "the concept of a mental state is primarily the concept of a state of the person apt for bringing about a certain sort of behaviour" (Armstrong 1968, 82). But he goes on to qualify that statement by saying that some kinds of "mental states will turn out to stand in still different causal relations to the behaviour which constitutes their 'expression'. In some cases, indeed, it will emerge that certain sorts of mental states can only be described in terms of their resemblance to other mental states that stand in causal relations to behaviour" (Ibid, 83). Other analytic functionalists stray even farther from their behaviorist predecessors. In listing differences between the two views, Lewis notes that analytic functionalism and not logical behaviorism "allows us to include other experiences among the typical causes and effects by which an experience is defined" (Lewis 1966, 21). Schwitzgebel, in listing the sort of dispositions (which he takes to be essentially the output side of functional roles)[58] lists *phenomenal* and *cognitive* dispositions—"dispositions to have certain sorts of conscious experiences" and "dispositions to enter mental states that are not wholly characterizable phenomenally, such as dispositions to draw conclusions" respectively (Schwitzgebel 2002, 252)—as *sui generis* categories of output alongside behavioral dispositions, giving no pride of place whatsoever to the latter.

The upshot of why Block's criticism is not fatal to analytic functionalism is that we need to think of it not just as a theory that relates mental states to the external but, to each other, and the latter has at least equal importance.

Part Two: Analytic Functionalism and Liberalism

Block (1978) charges all varieties of functionalism with being either too 'liberal' or too 'chauvinist.' A theory of mental states is too liberal if it assigns mentality to beings that do not in fact have it. It is too chauvinist if it denies the mentality of beings that do in fact have it. Seen another way: liberalism is about setting the bar for sufficient conditions for mentality too low; chauvinism is about setting the bar for necessary conditions for mentality too high. Block argues that analytic functionalism is guilty of liberalism, whereas psychofunctionalism is guilty of chauvinism. I will discuss the first charge

[58] "One might think of a dispositionally specifiable state as a state of an object, e.g., a brain, apt to bring about specified effects under specified conditions, and of a functionally specifiable state as a state of an object apt to bring about specified effects under specified conditions and to be produced by specified causes." (Schwitzgebel 2002, 257).

here—which I deny—and discuss the second charge in the next chapter (with which I will agree).

Preliminaries

A few points of clarification are needed before proceeding to the arguments. First, there are two *versions* of the charge of liberalism. One version—and the one I will focus on here—is about mentality generally. It is about whether a creature can have any kind of mental state whatsoever. The other version is specifically about *phenomenal* states; whether a creature can have states regarding what it is like to be a certain way. Block's own position, more precisely stated, is that while analytic functionalism is too liberal and psychofunctionalism too chauvinist with respect to mentality generally, both are too liberal with respect to mentality *qua* phenomenality. For, as he argues, it is not clear that having identical functional states to beings that have phenomenal experiences (like us) under *any* specification is sufficient for phenomenality. Indeed, this sort of argument has motivated some of the most oft-discussed objections to functionalist phsyicalisms (Jackson 1982; Chalmers 1996). However, as Lycan (1979) points out, Block's rationale is simply begging the question. For it denies precisely what all functionalists assert: that functional equivalence, of some variety, is sufficient for all types of mentality. I do not want to here settle the issue of the relationship between functionalism and phenomenal experiences. I will therefore restrict my discussion to the charges of liberalism and chauvinism with respect to mentality *simpliciter*; it is fodder for discussion enough.

Block specifies the conditions under which analytic functionalism and psychofunctionalism assign mentality in terms of functional equivalence. I should thus say a bit about this concept. Block discusses three kinds of functional equivalence, which exist in a sort of hierarchy. The most lax is input-output equivalence, in which two systems are equivalent if they produce the same output (behavior) in light of the same input. This sort of equivalence says nothing about what is going on inside the system or organism. The most strict is psychofunctional equivalence, wherein a psychofunctional Lewis-Ramsey sentence must be true of both systems or organisms for them to be equivalent. And in between is analytic-functional equivalence, wherein an analytic-functional Lewis-Ramsey sentence must be true of both systems or organisms for them to be equivalent.

I will call two systems that are equivalent at any level to be ‘functional equivalents’ at that level. And I classify this hierarchy in terms of levels of abstraction: the level of behavioral equivalence is the ‘highest’ or most abstract, the level of analytic-functional equivalence lower or less abstract, and the level of psychofunctional equivalence the lowest or least abstract.

The liberalism argument against analytic functionalism is that some systems which are analytic-functionally equivalent to humans lack mentality and should not be assigned it. The chauvinism argument against psychofunctionalism is that some systems which are not psychofunctionally equivalent to humans have mentality and should not be denied it.

Homunculi-Heads

Block’s argument that analytic functionalism is guilty of liberalism is by far the more famous of the two (even though, by my lights, it fails while his chauvinism argument succeeds). Block gives two examples of entities he takes to be (potentially) analytic-functionally equivalent to humans that by his lights, intuitively lack mentality. In his words:

Imagine a body externally like a human body, say yours, but internally quite different. The neurons from sensory organs are connected to a bank of lights in a hollow cavity in the head. A set of buttons connects to the motor-output neurons. Inside the cavity resides a group of little men. Each has a very simple task: to implement a "square" of a reasonably adequate machine table that describes you. On one wall is a bulletin board on which is posted a state card, i.e., a card that bears a symbol designating one of the states specified in the machine table. Here is what the little men do: Suppose the posted card has a 'G' on it. This alerts the little men who implement G squares-'G-men' they call themselves. Suppose the light representing input I 17 goes on. One of the G-men has the following as his sole task: when the card reads 'G' and the I 17 light goes on, he presses output button 0191 and changes the state card to 'M '. This G-man is called upon to exercise his task only rarely. In spite of the low level of intelligence required of each little man, the system as a whole manages to simulate you because the functional organization they have been trained to realize is yours (Block 1978, 278).

> Suppose we convert the government of China to functionalism, and we convince its officials that it would enormously enhance their international prestige to

> realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants.) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. We replace the little men with a radio transmitter and receiver connected to the input and output neurons. Instead of a bulletin board, we arrange to have letters displayed on a series of satellites placed so that they can be seen from anywhere in China. Surely such a system is not physically impossible (Ibid, 279).

There are two points worth noting. First, what the two examples have in common is that they are systems whose basic units are themselves minded beings (homunculi) who each play very simple roles. Hence the name 'homunculi-heads.'

Second, the conception of functionalism involved in both examples is not the one I have been discussing as standard. The examples are set up as equivalent to Turing Machines, which realize one state at a time. Each 'state' specifies a given output and possible transition to another state depending on what is inputted to the machine. They are thus functional systems, i.e. systems definable in terms of inputs, outputs, and relations to other states. In contrast, the sort of functionalism I have been describing, where functional roles are defined via a Lewis-Ramsey sentence, is best understood as 'causal role functionalism' since the relations specified in the Lewis-Ramsey sentence are causal in nature.[59] CR-functionalism does not require that there only be one state at a time, for it

[59] Polger (2004) distinguishes between causal-role functionalism and 'theoretical functionalism,' identifying the Lewis-Ramsification method with the latter. His point is that the nature of a Lewis-Ramsey sentence does not logically entail that the relations between variables are causal in nature. While Polger is

does not involve a single machine of which states may be predicated. Conversely, Turing Machines do not imply that the relations between states are causal; Turing Machines are purely formal devices. Historically, it was Putnam (1967) who introduced TM functionalism whereas Lewis (1966; 1972; 1980) and Armstrong (1968) introduced causal-role functionalism.

I think it is fair to say that intuitions in Block's examples speak against either homunculi-head having a mind. However, we should take care to ensure that this intuition is not polluted by their being TM-functional systems as opposed to systems whose individual states are defined in terms of causal/functional roles. For, as Block himself has noted (Block and Fodor 1972), TM-functionalism is untenable for a variety of reasons that are unique to it. For instance, Turing Machines are only in a single state at a time, yet people clearly have many beliefs and desires active ('occurrent') at any given moment. One can attempt to avoid this criticism by identifying a given TM state with the entire set of mental states a person has active at a time, however, this solution won't work for reasons Block himself details: "This version of Functionalism … has no resources for appropriately handling the content relations among mental states, e.g., the relation between the belief that P and the belief that(P or Q)" (Block 1978, 283).

Let us thus redescribe the homunculi-head example for causal role functionalism. Take the Lewis-Ramsey sentence that analytic functionalism will say is true (of minded beings as such). This sentence will likely have millions of variables, each of which is related to

surely right, his point is orthogonal to the current discussion: I am assuming the relations among the variables of the Lewis-Ramsey sentence *are* causal in nature for the theories discussed here.

hundreds or thousands (or millions) of other variables via causal roles. By causal role relations, I just mean things like: X causes Y, Y is caused by X, or X requires the presence of Z to cause Y. Now have a person, or several people in an entire nation 'play the role' of each variable. Take the people playing the role of J. Maybe they can use a system of relay lights (like ships did before radio communication) to alert other people, playing variables related to theirs via the Lewis-Ramsey sentence, when the variable they are playing is 'active.'[60] Say, J causes K. Then those other people, playing K, would have an instruction manual, telling them that when the J squad activates their relay lights, they, the K squad, should do the same. And when the K squad activates their lights, it will let further relevant people, perhaps those playing L, know that K is active, etc.

In theory, a nation with enough people could perfectly realize the Lewis-Ramsey sentence. Now suppose that somewhere else, there is a perfectly normal human body, except its brain has been replaced by a computer that is programmed with the Lewis-Ramsey sentence the billion people over in, say, India, are realizing. And suppose it is somehow monitoring their activity (perhaps the lights are really bright and visible from space). Finally, suppose that, according to the Lewis-Ramsey sentence of analytic functionalism, X is the belief that there is beer in the nearest fridge, Y is the desire for beer, and Z is the behavior of going to the fridge. The Lewis-Ramsey sentence will thus have a line that says [X, in the presence of Y, leads to Z]. As soon as the X squad and the

[60] I had considered, in crafting this example, people using smartphones or tablets to communicate with each other. But then it could be objected that the network of phones is really realizing the Lewis-Ramsey sentence, not the people operating them, as the information merely needs to pass through the network of smartphones in order to activate K based on J, L based on K, etc. With relay lights, the informational connection must go through the minds of the operators.

Y squad flash their lights, the computer moves the body it inhabits in a way that perfectly matches what you or I would do that could be described as going to the fridge and getting beer.

The analogue to Block's original argument is that this system, too, intuitively lacks mentality. And if it does, then analytic functionalism is too liberal for assigning it mentality—the relevant Lewis-Ramsey sentence is, after all, true of it.

A Failed Response

One strategy of dealing with homunculi-heads is to say that they are distinct from humans and other (intuitively) minded organisms in salient ways. I am skeptical of this line of thinking generally. The most common way it has been cashed out was suggested by Putnam (1967), who claimed that minded beings could not 'nest,' i.e. have other minded beings as their components (see Kammerer 2015 for elucidation and defense). However, Block points out (see also Schwitzgebel 2015; 2016) why this response won't do. Clearly, there are *some* minded beings who contain other minded beings: otherwise we would have to say pregnant women, the moment before birth, would not have minds, even though they clearly did before becoming pregnant (and before the fetus developed its brain sufficiently), and will again *the very moment* they give birth.

On Putnam's behalf, Block reformulates the anti-nesting argument to the claim that: "a painfeeling organism has a certain functional organization and that it has no parts which (1) themselves possess that sort of functional organization and also (2) play a crucial role in giving the whole system its functional organization" (Block 1978, 291). Block argues that this won’t do either: there could be a part of our universe where there are sentient beings the size of atoms over in our neck of the celestial woods. Suppose that these atomic men decide to function just like our atoms, and some of them end up replacing some atoms in our brains. Surely, Block says, this would not mean we cease to have minds!

By making this last point Block concedes something important: that at least *some* homunculi-headed beings can be minded. Further, Block asserts that a homunculi-head that is *psychofunctionally* equivalent to us would have mentality. He argues “Since a Psychofunctional simulation of you would be Psychofunctionally equivalent to you, a reasonably adequate psychological theory true of you would be true of it … What better reason could there be to attribute to it … mental states? … In the face of such a good reason for attributing mental states to it, prima facie doubts about whether it has those aspects of mentality which are in the domain of psychology should be rejected” (Block 1978, 301).

Block explains his reasoning for thinking complete physical equivalence (when the homunculi replace the atoms or molecules in our brains with functionally identical ones) and psychofunctional equivalence stand out. “What seems important is how the mentality of the [homuncular] parts contributes to the functioning of the whole” (Ibid, 292). The homunculi in an atom/molecule-level equivalent or a psychofunctional equivalent are functionally indiscernible at a fine enough level for the fact that they are homunculi not to make a difference to the mentality of the overall entity.

Persuasive Negative Responses

I think there are two responses that do have some teeth against Block’s argument. One is due to Lycan (1981); the other is my own. But these responses are limited: what they show is that Block’s argument does not in fact prove that an analytically-functional equivalent homunculi head lacks a mind. What they do not show is that such a being has a mind.

The key point to carry over from the previous section is that Block wants to draw a line: given equivalence at or below a certain level of abstraction, we should override our *prima facie* intuitions that a system lacks mentality and grant it to said system on the basis of this kind of equivalence. But for every system that is only equivalent to you at a more abstract level—a level above this ‘magic,’ mentality-endowing line—we can let our *prima facie* intuitions carry weight. His argument that analytic-functional equivalence is insufficient for mentality boils down to the claims that i) it is unintuitive to assign them mentality, and ii) analytic-functional equivalence is above (more abstract than equivalence at) the magic line.

Lycan (1981) strongly rebuts i). He writes:

> Suppose that you were a little, tiny person-say, just ten times the size of a smallish molecule. And suppose that you were located somewhere within Ned Block's brain, perhaps standing somewhere in his left occipital lobe. What would you see? It would seem to you that you were standing in the middle of a vast and largely empty space. Occasionally a molecule (looking something like a cluster of basketballs) would whiz by at a terrific rate; sometimes you would see two or more of these clusters collide and rebound (Lycan 1981, 38).

Lycan suggests that from this perspective, it would seem astounding if the physical activity you were witnessing constituted mentality. But clearly it does. It would seem absurd “[p]resumably because you would be too small to see the forest for the trees” (Ibid). Lycan then suggests that “[s]imilarly, Block is too small to see the Chinese mainland's inputs and outputs, respectively, as psychological stimuli and behavior to which he could relate” (Ibid), and, further, “[i]n the case of the homunculi-head, I suggest, the same Gestalt failure obtains even though Block is not smaller than the organism; Block's attention is focused on the hectic activities of the little men, and so he is seeing the homunculi-head as if through a microscope” (Ibid, 39).

Lycan’s idea is that it is never intuitive that some physical system constitutes mentality when looking at the inner workings as individual physical processes. Block finds Lycan’s argument persuasive, conceding “I think that the Lycan … point does genuinely alter one's intuitions” (Block 1981, 42n30). At the very least, we can conclude from Lycan’s argument that intuitions alone are not going to settle whether (merely) analytical functionally equivalent homunculi heads have or lack mentality.

But Block could reformulate his argument to get around this concern. The ‘magic line’ is doing the real conceptual work here: after all, it overrides our intuitions about what does or does not have mentality. If something is below (less abstract than equivalents at) this line, we should still refrain from attributing to it mentality, even if our intuitions are silent on the matter. And Block does maintain something like this view. Even granting Lycan’s rebuttal to his intuitions, he contends that a system which is merely behaviorally equivalent to a minded being is insufficient for mentality. Consequently, *some* internal difference in information processing style makes a difference for mentality.

While this response allows Block to stay in the conversation, and potentially claim that the analytic-functionally equivalent homunculi-head is not minded, it represents an important concession from his initial argument. Specifically, it implies that whether or not a being is a homunculi-head is irrelevant for its mindedness. The entire issue is now about the complexity of its inner structure. And while there is still a potential case to be made for functional equivalents at some level of abstraction being non-minded, the argument has lost the aspect that I take pulled many readers in at the beginning: that it was problematic for any sort of functionalism to imply that any kind of homunculi-head had a mind of its own. Nevertheless, we move on.

Block goes on (in his [1981]) to give an example—what has come to be known as a 'Blockhead'—of a system that can act in conversation as a normal human would, but does so in virtue of having a (very, *very*) long list of acceptable responses to any conversational 'move' and simply picking one from the least each time its interlocutor speaks. He—rightly, in my view—argues that such a system would not have mentality.

On this last point, I agree with Block. The Blockhead is below the magic line. The field of animal cognition has yielded several attempts at demarcating a level of internal complexity that separates minded from non-minded organisms. From the very strong, such as Davidson's (1982) contention that having a language is a necessary condition for mindedness, to the weaker, such as Saidel's (2009) contention that goal-directed learning is sufficient for mindedness, the Blockhead fails them all.

Crucially, however, the Blockhead *is not analytic-functionally equivalent to us!* It is merely behaviorally (input-output) equivalent. For analytic functionalism includes a great deal of internal structure—about how states and processes relate to each other to alter states (like revising beliefs), generate new states (such as by completing inferences), and so on. (Indeed, the lack of attention to this rich inner structure is one of the flaws with Block's argument from paralytics).

Let us take stock. There are two crucial points here. First, there is some line, which I have been calling the 'magic line,' for which beings that are equivalent to us at a certain level of abstract description—or above that level—have minds even if they are homunculi-headed. Second, given the only example that we can conclusively say is below the magic line, said line is (much) lower than analytic-functional equivalence. Given these two points, Block's homunculi-headed argument does not give us a positive reason to think that analytic functionalism is too liberal.

A Positive Argument

I want to go farther, though. I want to conclusively say that an analytic-functional equivalent homunculi-head *has* a mind. I will now attempt to make an argument to that end. What I will argue is two beings in an example Block discusses and explicitly (and in my view, correctly) claims are both minded are equivalents at the degree of abstraction at which analytic functionalism operates. Specifically, if some being is analytic-functionally equivalent to a certain human, then it must make the same inferences she does. And, because Block takes it to be chauvinistic to say the beings equivalent to humans at this level of abstraction lack mentality, by his own lights, analytic-functional equivalence to a human being implies mentality.

Where Block errs is in attributing insufficient richness to analytic functionalism.

Block argues that only psychofunctional—and not analytic-functional—equivalence implies equivalence in information-processing. He notes that if a class of systems "are construed as [analytic] Functional (rather than Psychofunctional) simulations, they need not be things to which psychological (*information-processing*) theories true of us apply" (Block 1978, 292, italics added). But he goes further than saying that analytic functionalism does not imply information-processing equivalence. He says—more than once—that it doesn't imply *anything like* information-processing equivalence: an analytic-functional equivalent system "need not be anything like you psychologically (that is, its information processing need not be *remotely* like yours)" (Ibid 296; italics added); and that "[analytic] Functional simulations need not have … psychological (information-processing) … *anything like* ours" (Ibid 301, italics added). Block's rationale is that information-processing is something only found in a scientific psychology. As he says, "if current theories of psychological processes are correct in adverting to storage mechanisms, list searchers, item comparators, and so forth, Psychofunctionalism [but not analytic functionalism] will identify mental states with [inputs and outputs to] causal structures that involve storage, comparing, and searching processes as well as inputs, outputs … [to] other mental states" (Ibid 274).

I think Block is right that there is *a sense in which* information-processing is proprietary to scientific psychology, and thus psychofunctionalism. But information processing *simpliciter* is not.

The issue, however, is that there are two different ways of understanding information processing, and only one is proprietarily psychofunctional. We might understand information processing in terms of information to be processed, as Dretske (1981) does. Or we might speak of actual systems processing information, which do so in some sort of language or other medium—a formal system—that represents the information processed. There is a fairly large difference between the two.

Block (1981) gives an example of a hypothetical species of Martian that are (at least) behaviorally, but not psychofunctionally, equivalent to humans. Suppose both species make the same inferences, and in the same language (say, English). But their cognitive architectures use different 'strategies:' "One strategy would be to represent the information in the machine in English, and to formulate a set of inference rules that operate on English sentences. Another strategy would be to formulate a procedure for translating English into an artificial language whose sentences wear their logical forms on their faces … Suppose that the Martian and human psychologists agree that Martians and humans differ as if they were the products of a whole series of engineering decisions that differ along the lines illustrated" (Block 1981, 6). Block then asks: "[s]hould we conclude that the Martians are not intelligent after all? Obviously not! That would be crude human chauvinism" (Ibid).[61]

I claim that the human and Martian are equivalent *qua* information-processing in the Dretskean sense. And I further argue that not only are the two analytic-functionally equivalent, but the specific (least abstract) level at which they are equivalent is the level of analytic-functional equivalence. Thus, they are equivalent at the level of information-processing proprietary to analytic functionalism. Further, given that i) an entity's being a homunculi-head does not speak to whether or not it is minded, and ii) Block's own admission that both equivalents in this scenario are minded, an analytic-functionally equivalent homunculi-head is minded.

I now turn to the work of David Marr (1982) for a way of distinguishing different ways of being equivalent *qua* information-processing. On my reading of Marr, his distinction between information at the computational and algorithmic levels perfect maps onto the respective senses of information-processing proprietary to analytic functionalism and psychofunctionalism. This interpretation of David Marr is nonstandard, but, I think,

[61] Block gives another example to make the same point in his (1978). There he says, "the difference [between humans and Martians] can be described as follows. Think of humans and Martians as if they were products of conscious. design. In any such design project, there will be various options. Some capacities can be built in (innate), others learned. The brain can be designed to accomplish tasks using as much memory capacity as necessary in order to minimize use of computation capacity; or, on the other hand, the designer could choose to conserve memory space and rely mainly on computation capacity. Inferences can be accomplished by systems which use a few axioms and many rules of inference, or, on the other hand, few rules and many axioms. Now imagine that what Martian and Earthian psychologists find when they compare notes is that Martians and Earthians differ as if they were the end products of maximally different design choices (compatible with rough Functional equivalence in adults)" (Block 1978, 310-11). My sense is that Block changed his example because the later one is more clear and persuasive. But if, for any reason, one is not persuaded by the example cited in the main text, they may substitute this—or any similar example they can imagine based on the principles more abstractly articulated in this passage—to make the same point.

correct. It is from my own reading of Marr as well as a recent paper by Ritchie (2019) that independently came to the same conclusions as I did.[62] But the way in which the traditional interpretation of Marr is incorrect will, I think be illuminating and show why Block is underestimating the extent to which information-processing plays a role in analytic-functional equivalence.

David Marr and the level of Computational *Theory*

Marr famously distinguishes between a computational level and an algorithmic level of analysis.[63] A standard account of Marr attributes to the computational level of analysis the inputs and outputs of a function, leaving a 'black box' in between. The top or computational level "describes the task which the psychological system accomplishes" (Griffiths 1997, 220); it "specifies what processes do, without specifying how they do it" (Barrett 2009, 332n8), and it is "a purely input/out[put] level, at which we simply describe the input and output of [a] … system" (Rey 1997, 180). Whereas the middle or algorithmic level[64] "describes how the system processes information in order to accomplish this task" (Griffiths 1997, 220); "provides a description of the logical steps that are needed to implement the computational level" (Barrett 2009 332n8), and it

[62] Ritchie (personal communication) and I agree that, on the one hand, it is noteworthy how close our independent readings of Marr turned out to be. On the other hand, we both agree that it is somewhat surprising and more than a bit disappointing that so many excellent scholars of Marr have missed our reading. For our understanding differs from earlier interpretations mainly in that it relies on the later chapters as much as the initial ones. Many interlocuters of Marr seem only to cite his first chapter, and none (save Ritchie) that I know of cite the final two.

[63] At times, Marr calls the algorithmic level the 'representational' level; I will stick to the label 'algorithmic' as—for reasons we will see—it will make explication much more clear and simple.

[64] The bottom is the 'implementation' level, which deals with neural mechanisms

“describes the procedures … that are responsible for computing the output from the input” (Rey 1997, 180).

These readings are basing their interpretation on Marr’s definitional gloss of the computational level in his introductory chapter: “What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?” (Marr 1982, 25). Philosophers have thought that he meant something like: early vision is best decomposed into functions that start with edge detection, go on to 2D image creation, etc.

The first function computed by the visual system, according to Marr (1982, 51) is to locate intensity changes, which represent edges, in an image. It takes as input the raw data from the retina and computes where the edges of objects in the visual field are. In short, it is an edge detection function. On the traditional understanding of the computational level, I have just given a complete description of the first function: it is an edge detection function, taking raw retinal data as input and providing representations of edges of objects as output. But it is not.

For something is rotten in the town of Cambridge. As Egan (1992) notes, Marr himself says, a ‘LaPlace-Gaussian’ mathematical function is “from a computational [level] point of view … a precise specification of what the retina does” (Marr 1982, 337). Rey finds this claim of Marr’s “puzzling … surely the Laplace-Gaussian is closer to a ‘level-2,’ ‘algorithmic’ characterization: although it doesn’t specify an algorithm, it does provide the mathematical characterization of … the algorithm, one step closer than ‘edge’ to a level-2 ‘choice of representation for the input and output” (Rey 2003, 122). While some previous work (Egan 1992; Shagrir 2010) has been sensitive to this distinction, I believe Marr’s remarks have not been properly appreciated until Ritchie (2019).

According to Ritchie, each level has a ‘representation’ and a ‘process’ component (see also Marr’s fig 6-1, p332 in his [1982]). The computational level is best understood as a level of “the computational *theory*” of a function (Ritchie 2019, 1080). A computational theory is a conceptual account of what strategy—what *kind* of algorithm is the best strategy for performing a function. The computational level also contains a justification of why that strategy is the right one. The level contains both the blueprint and the justification for each of ‘representation’ and ‘process.’ At the computational level, these components correspond to: what information needs to be represented, and why? And what type of algorithm—in terms of Rey’s ‘mathematical characterization’—is best to use, and why? Together, Ritchie (Ibid, 1081) claims, these components give a *blueprint* for an information processing-device.

Understood this way, Marr is saying that a particular mathematical function, the LaPlace-Gaussian, is the right process to perform edge detection and thus transform raw retinal data into information about the location of object edges in a visual image. And Marr indeed goes to great length to explain why this is the case! Moreover, he does so only after explaining what kinds of information we should include in the input and output of that function. I discuss these details, as well as a more general case for Ritchie’s/my interpretation of Marr in the following ‘Sub-Appendix.’

The algorithmic level also contains the representation and process components. But here the representation component is a specification of what representational *format* (what ‘language’ in Fodor’s terminology). And the process component is a specification of what specific *values* should be used in the algorithm specified at the computational level (e.g. what graphical resolution should the Gaussian function use?).

Perhaps the reason philosophers have missed this subtle distinction (Marr acknowledges “it is probably one of the most difficult ideas … to grasp” [Marr 1982, 330]) is that Marr’s example in his conceptual/philosophical introductory chapter for distinguishing between the computational theory level and the algorithmic level is addition. For addition, there is only one possible algorithm to use for the ‘blueprint’—an addition algorithm. There are no values to specify. We don’t do addition at X or Y resolution; we don’t have constants in a complex algorithm to set. Thus, on the process side of things, the distinction between the computational level blueprint or mathematical specification of function and algorithmic level specific values to be used in the function goes easily unnoticed.

But even in this simple example, there is an overlooked difference on the representation side of things. For while the algorithm is implied by the blueprint—and thus the process aspect is oversimplified—the representational format is not implied by the ‘computational theory’ idea that it is numbers that need to be represented. “[H]ow information is represented can greatly affect how easy it is to do different things with it. This is evident even from our numbers example: it is easy to add, to subtract, and even to multiply if the Arabic or binary representations are used, but it is not at all easy to do these things—especially multiplication—with Roman numerals" (Ibid, 21). That is, there are multiple representational formats for numbers, and, even once the 'computational theory' of addition is settled by the very fact that the input and output must be numbers.

How Marr discusses making the choice of representational formats is noteworthy, and will return us to the actual topic at hand:

> An analogous problem [to the one just described] faces computer engineers today. Electronic technology is much more suited to a binary number system than to the conventional base 10 system, yet humans supply their data and require the results in base 10. The design decision facing the engineer, therefore, is, Should [sic] one pay the cost of conversion into base 2, carry out the arithmetic in a binary representation, and then convert it back into decimal numbers on output; or should one sacrifice efficiency of circuitry to carry out operations directly in a decimal representation? (Ibid).

Compare this to Block's discussion of humans and Martians (I here reprint the quotation):

> One strategy would be to represent the information in the machine in English, and to formulate a set of inference rules that operate on English sentences. Another strategy would be to formulate a procedure for translating English into an artificial language whose sentences wear their logical forms on their faces … Suppose that the Martian and human psychologists agree that Martians and humans differ as if they were the products of a whole series of engineering decisions that differ along the lines illustrated (Block 1981, 6).

These two passages are discussing *exactly the same issue*. They are discussing how to go from the stage of having a blueprint or 'computational theory' of, in one case mathematical functions and in the other case inference, and figuring out what algorithms (and possibly additional processes) should be used to carry out that strategy.

I think we are now in a position to make several conclusions. First, there are two senses, as delineated by Marr, in which we might discuss information processing: in terms of the computational theory of an information-processing task, abstracting away from the representational medium and specific, 'programmable' algorithm, and in (algorithmic level) terms of just that medium and algorithm. Second, Armstrong cares about information processing in the former sense: he does not care about how one instantiates or implements the sensitivity of intention-states to information from perception, he just cares that the 'strategy' of intention is to be sensitive to that information.[65] Third, the mistake Block is making in his characterization of analytic functionalism is the very same mistake that Rey, Griffiths, and Barrett make in misunderstanding Marr's computational level. Block is right that there is a sense in which analytic functionalism does not require equivalence of information processing. It is the sense Marr discusses at the middle or algorithmic level. But it doesn't follow that analytic functionalism allows radically different versions of information processing. Rather, analytic-functional equivalence allows for differences at the algorithmic level but not at the level of computational theory. Block's mistake is understandable:[66] without Marr's conception of computational

[65] Note that this argument has the form of an inference to the best explanation. The only way of describing how one can care about an information-processing task without caring out the programmable representational medium and algorithm—the 'only game in town'—to borrow a phrase from Fodor (1987)—is Marr's notion of 'computational theory.'

[66] In fairness to Block, Marr hadn't yet written his analysis while Block was writing the two papers of his I've focused on in this section. And even after Marr did write it, he was consistently mischaracterized (indeed, he at times still is: see Rescorla, 2020). But with the right understanding of Marr's notion of computational theory in hand, we can see how there is a sense—consistent with, and, indeed, required to

theory, *there is no way to understand information-processing tasks* qua *information processing-tasks without adverting to specific, programmable, representations and algorithms.* And because *that* way of understanding information tasks is clearly the proprietary territory of psychofunctionalism—as Block's Martians demonstrate—it can't be part of analytic functionalism.

The Upshot

Block was quite right about the basis on which we assign mentality. It has to do with the complexity of the parts and the way they interact. And it seems entirely plausible to me that what it is to be minded is closely linked to the idea of processing information. What this section demonstrates is that analytic functionalism has a rich notion of information-processing among its explanatory resources. The mere idea that thought is computational and representational does not itself put analytic functionalism to rest as a viable theory of the mind. While psychofunctionalism can certainly investigate information processing in details unavailable to analytic functionalism, including the representational format of thought, these details seem to go beyond what is essential to being minded.

Appendix 4 / Sub-Appendix, or, More on Marr

Here I will discuss the case for Ritchie's and my unorthodox interpretation of Marr. The first aspect is his discussion of the LaPlace-Gaussian function, and the lengths he goes

make sense of, Armstrong's discussion of intention—in which analytic-functional equivalence very much requires information-processing equivalence.

through to explain why it is a good computational ‘blueprint’ for edge detection. This is consistent with our reading of Marr, but completely superfluous on the traditional reading. Second, I will discuss other passages of Marr that only make sense on our reading.

LaPlace-Gaussian

In short, what we want to do to detect edges is find high magnitude changes in light intensity; they indicate boundaries of objects. Marr argues that the best way to do this mathematically is via a LaPlace operator: it is a mathematical function to map intensity. What Marr says is especially good about the operator is that it is a second-order derivative. First order derivatives, Marr notes, “could be used, in which case one would subsequently have to search for their peaks or troughs … However, the disadvantage … is that they … involve an orientation” (Ibid., 56-57). The first order derivative of an intensity mapping will show a ‘peak’ (when the algorithm ‘moves’ to a much brighter part of the visual field) or a ‘trough’ (when it moves to a much darker part) where there is a great change in intensity, but edge detection doesn’t depend on whether the object occludes more or less light than its surroundings. Thus the orientation information is superfluous, and it would be best to find a mathematical process that could ignore it. Using the second derivative turns both peaks and troughs alike into ‘zero-crossings,’ a point (or small space) on a graph where the value goes from positive to negative or vice-versa. This allows the function to leave out the extraneous data.

But not all intensity changes in the visual field map onto object boundaries. Sometimes, there are intensity changes *within* an object. To get rid of those, we want to use a 'blurring' function that only allows intensity changes that persist over a good amount of the visual field to make it through the blurring process. The intensity changes that take up more space are, generally speaking, going to correspond to the boundaries of an object rather than shading within the object because things within the object are smaller than the object itself. "The Gaussian part … blurs the image, effectively wiping out all structures much smaller than [the object being resolved] … The reason why one chooses the Gaussian for this purpose … is that the Gaussian distribution has the desirable characteristic of being smooth and localized in both the spatial and frequency domains … And the reason, in turn, why this should be a desirable property of our blurring function is that if the blurring is as smooth as possible … it is least likely to introduce any changes that were not present in the original image" (Marr 1982, 56).

To make a long story short: Marr goes to great detail to explain why the LaPlace-Gaussian function is the right one for edge detection. His reasons for doing so are that he thinks analysis of information processing takes two distinct forms. Of course, there is the analysis of what representational medium, and what specific algorithm (LaPlace-Gaussian doesn't specify a precise algorithm, e.g. it doesn't specify the resolution of the blurring) need to be used. But Marr thinks there is a distinct step of conceptual analysis where we figure out a 'blueprint' for performing the function before we get to the technical details.

Other Support

Leaving the domain of edge detection, Marr’s conceptual/philosophical discussion repeatedly emphasizes the conceptual side, the ‘computational theory,’ of analyzing information-processing. What Marr meant by “what is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it is carried out” is: what is the task (edge detection), what is our strategy for the task (a LaPlace-Gaussian type of computation), and why is that blueprint the one we should use (the explanation I gave).

Marr makes several remarks that are quite puzzling if we read him in any other light but straightforward if we read him in this one. “For far too long, a heuristic program [specific algorithm] of carrying out some task was held to be a theory of that task, and the distinction between what a program did [the blueprint] and how it did it was not taken seriously enough” (Ibid, 28). This makes no sense if ‘what a program did’ was the function itself—edge detection. Philosophers and psychologists were quite aware that edge detection and an algorithm for edge detection were different things! Marr’s point is that specifying a general blueprint-level strategy (LaPlace-Gaussian) and an actual algorithm which instantiates a LaPlace-Gaussian are not the same thing. As he reiterates near the end of his work, to understand vision, “It is not enough to be able to describe the responses of single cells [implementation-level analysis] … Nor is it enough even to be able to write computer programs that perform approximately in the desired way. One has to do all of these things at once and also be very aware of the additional level of explanation that I have called the level of computational theory” (Ibid., 329-330). If the level of computational theory is just a specification of the inputs and outputs of a device, this comment is likewise puzzling. But it makes sense if computational theory is *a distinct way of analyzing information-processing*.

That this is Marr's meaning can be further demonstrated by his comments on Gibson's (1966) ecological theory of perception. Per Marr, Gibson was "perhaps the nearest anyone came to the level of computational theory … Gibson's important contribution was to take the debate away from the philosophical considerations of sense data and the affective qualities of sensation and to note instead that the important thing about the senses is that they are channels for perception of the real world outside" (Ibid, 29). I am not sure how to interpret these remarks on the traditional reading of Marr. But on my and Ritchie's reading, it is that Gibson figured out that the 'blueprint' for perception didn't involve a sense data representation plus a hypothesis confirmation function, *a la* von Helmholz but the process of recovery of information from the environment. And, for Marr, it is recovery of *information*: "Gibson's … fatal shortcoming … results from a failure to realize … the detection of physical invariants like image surfaces, is exactly and precisely an information processing problem" (Ibid, 30).