EVOLUTIONARY ARGUMENTS AND THE MIND-BODY PROBLEM

by

JOSEPH ANTHONY CORABI

A Dissertation submitted to the Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Brian McLaughlin

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2007

ABSTRACT OF THE DISSERTATION

Evolutionary Arguments and the Mind-Body Problem

By JOSEPH ANTHONY CORABI

Dissertation Director:
Brian McLaughlin

Imagine slicing your hand with a steak knife. Inevitably, this leads to a characteristic unpleasant sensation, and just as reliably, to a withdrawal of the wounded limb. But can this rather mundane fact—and other similar facts—shed any light on the mind-body problem or the issue of the role of experience in causing behavior? In my dissertation, I explore this issue head on, and in the process clarify and criticize the arguments of philosophers who have given an affirmative answer to this question—philosophers such as William James and Herbert Spencer. These arguments have coupled evidence like the above with the fact that human beings have evolved, in order to make the case that epiphenomenalism with respect to qualia is false.

My first task will be to formulate a rigorous version of a James-Spencer style argument, which will occupy us in Chapter 1. Chapter 2 is dedicated to answering a number of objections to the argument, in an effort to show that if there is a problem with it, this problem lies elsewhere. Chapter 3 explores alternative arguments in the spirit of the original one formulated in Chapter 1, and discusses any resulting effects on the plausibility of the conclusion. Chapter 4 is the capstone chapter of the dissertation. In it,

I discuss a crucial objection to all arguments in the spirit of that given in the first chapter—namely, that physicalism has analogous flaws to epiphenomenalism where accommodating the relevant evidence is concerned. My conclusions in this final chapter are twofold. First, that even if there is no fatal flaw in the general strategy the evolutionary argument employs, it works against all forms of dualism, not just epiphenomenalism. And second, the accusations made in the objection are correct; physicalism suffers from problems analogous to those faced by epiphenomenalism (and, indeed, interactionism as well).

Although the primary findings of the dissertation are negative, there are many lessons we can take from them along the way. Most prominent among them is an improved perspective on the appropriate roles of empirical findings and armchair philosophical theorizing in debate over the mind-body problem.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to everyone who helped me in one way or another to complete this project. First, I would like to thank my graduate student friends who were kind enough to offer their input on my arguments in conversation and in some cases even read drafts of the material when it was at an inchoate stage. Regardless of this dissertation's present merits, it would have taken a far worse form without their help. I would like to thank especially Geoff Anders, Heather Demarest, Kate Devitt, Jeff Glick, Gabe Greenberg, Jonathan Ichikawa, Michael Johnson, Dan Kelly, Danny Korman, David Manley, Jenny Nado, Iris Oved, Doug Parvin, Angel Pinillos, Adam Sennet, Andrew Sepielli, Adam Swenson, Jason Turner, Ryan Wasserman, Jonathan Weisberg, Dennis Whitcomb, and Stephanie Wykstra. I also profited greatly from conversations with and feedback from Frank Arntzenius, Bob Pasnau, Audre Brokes, Mark Crimmons, Sandy Goldberg, Colin McGinn, John Hawthorne, Bill Robinson, Ted Sider, Todd Moody, and Dave Chalmers. I owe a debt to the latter two especially, both of whom helped me to first appreciate the potential relevance of evolutionary evidence in the mind-body sphere.

Of course, I must also thank my committee members for their comments and other feedback: Barry Loewer, Achim Stephan, Dean Zimmerman, and my chair Brian McLaughlin. I am especially grateful to Dean and Brian, both of whom provided a great deal of philosophical and professional help even before the dissertation process had begun.

And finally, I would like to thank my family for all of their support. It is to them that this work is dedicated.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Introduction

A cynic might be inclined to think that the classical issues in metaphysics of mind are paradigmatic examples of permanent philosophical stalemate. According to the cynic's postcard characterization, business as usual has ruled for centuries. Although recent philosophers have brought the finely toothed combs and razorlike prowess of their blossoming trade, all they've really accomplished in mind-body debate since Descartes and Hume is sharpen some of the edges. All the basic moves are still the same, and the fundamental impasses still impasses. Everyone knows the shortcomings of all the general positions, and everyone knows why those shortcomings can't be used to decisively favor one of the options over others (with the exception of epiphenomenalism, which everyone knows is crazy). Of late, increased infatuation with this elusive thing called "naturalism" has led most participants to lean toward the side of physicalism or materialism, but that is really just a philosophical fad. At the end of the day, nobody really knows and nobody ever will know, because old fashioned, armchair philosophizing is just not a powerful enough tool to answer the questions that have to be answered, and no other tool is available.

While no doubt something like the above musings have gone through the heads of many an undergraduate sitting through Philosophy of Mind 101, the basic mindset is not restricted to them. One does not have to travel far in the philosophical world to find someone who has devoted her life to the study of philosophy with a more or less similar outlook.

It is hard to know where to begin in criticizing the cynic's understanding of the debate, as it includes numerous oversimplifications and outright falsehoods. To be sure,

though, there are also kernels of insight in the cynic's bitter analysis, otherwise intelligent people would not be attracted to it. Perhaps the best candidate for a spark of insight is in the cynic's dissatisfaction with purely *a priori* work in settling the debate. Although the cynic undoubtedly underestimates the amount of progress that has been made, there is also no question that centuries of conceptual work (including a barrage of such work in recent decades) have failed to yield a consensus view on the matter among reasonable participants in the discussion. It is understandable how someone could grow frustrated with the *status quo* in light of the situation.

Perhaps it was a similar frustration that led philosophers to begin looking for alternative methods of answering the classical questions in the metaphysics of mind, most especially questions like "is the mind something over and above the brain?," and "do mental entities have causal effects in the physical world, and if so, how?" Although it would be impossible to survey all the different alternative approaches philosophers have taken, the purpose of this work will be to explore one such alternative.

Born in the wake of Darwin's publication of *The Origin of Species* at the hands of eminent names like Herbert Spencer and William James, evolutionary arguments were developed in the hopes of at least narrowing the playing field among live potential answers to these difficult questions—in particular, to show that epiphenomenalism, which is repugnant to many but easy to generate an argument against for few, is severely undermined by empirical considerations.

Although my purpose will ultimately be a critical one—to show that in fact these arguments fail to establish what they aim to establish—there is good reason to pay attention to the results even if one was not enamored with the specific approach to begin

with. Even though the journey will not give us the ultimate answers we seek to the fundamental questions in metaphysics of mind, by carefully analyzing evolutionary arguments and coming to an understanding of their shortcomings, one can be led to a much sharper comprehension of the issues in play and their interrelations with one another. In particular, the process will help us to delineate and appreciate the respective roles of empirical work and *a priori* philosophical theorizing in getting to the bottom of the debate—to understand where we must live with the age old controversies and where there is fresh hope of transcending them.

As an aid to the reader, let me say a bit more about the general layout of the work before diving into the main body of it. Some of the discussions will of necessity get intricate at times, so it will be useful to provide an indication of the "big picture" to prevent missing the forest for the trees. In the first chapter, I will begin by introducing epiphenomenalism, and explaining why some philosophers have been attracted to it in spite of its almost universally recognized counterintuitiveness. In the process, I will also introduce physicalism. After the preliminaries are complete, I will launch into a formulation of a restricted version of the evolutionary argument. The formulation will be more precise than traditional versions, but will pit epiphenomenalism against only physicalism. The argument contends that, once we are aware that evolution has occurred, physicalism does a much better job of leading us to expect the correlations between phenomenology and distal stimuli that we actually find, and so physicalism is made more likely to be true in the light of that evidence. My reason for beginning by considering only this narrow version of the argument is that most of the objections that have to be dealt with are either specific to epiphenomenalism, or focus on the abstract, general

dialectical strategy. Consequently, introducing the third major player into the debate at this stage—interactionism—would complicate the discussion unnecessarily and require straightforward but formulaic and tedious repetitions of various responses to objections and the like.

In the second chapter, I will address a number of objections to the argument presented in the first. Some of those objections will be relatively easy to refute, while others will take considerable time and effort to examine. By the end of the chapter, though, I hope to show that none of the objections (save one) is ultimately successful in undermining the argument. I will definitively reject some of the objections, while only sketching replies to the others (owing to their systematicity and the limit of space available). But these replies will nonetheless convincingly establish, I hope, that significantly more work would need to be done in developing the objections before they would constitute a serious threat. The one remaining objection, however, is serious and complicated enough that I wait until the work's final chapter to consider it in depth.

Chapter 3 contains a discussion of variations on the theme of the argument presented back in Chapter 1. In this chapter, I discuss three variations of the argument (variations in the kind of evidence considered, etc.) in an effort to learn whether they have any substantive effects on the overall prospects of the general strategy. I ultimately conclude that two of the three do not—they neither substantially strengthen nor substantially weaken the overall success of the basic line of argument—but that the other does (conditional on answering the lone remaining objection from the previous chapter). In fact, this other variation, which involves taking account of much more fine-grained

evidence than the original argument does, has the potential to significantly boost the pro-physicalist force of the main dialectical move.

In the fourth and final chapter, I return to the remaining objection from Chapter 2, and in the process introduce interactionism as an option in the debate. Once we have thoroughly examined this remaining objection and extended the original argument to include interactionism, we will be in a position to draw some ultimate conclusions.

First, we will see that traditional versions of this broader evolutionary argument, which have tended to group physicalism and interactionism together (as options confirmed by the evidence) and grouped epiphenomenalism by itself (as an option disconfirmed by the evidence), have made incorrect—or at least overly simplistic—inferences about how the general dialectical strategy should be employed, even assuming the remaining objection from Chapter 2 is faulty. More specifically, they have been mistaken in thinking that interactionism and physicalism were both supported by the evidence, and epiphenomenalism alone undermined by it.

Second, and perhaps more significantly, we will be in a position to see that the remaining objection from Chapter 2 does in fact undermine the entire central move of the argument. Consequently, the evidence typically adduced turns out to be useless in helping us to answer the metaphysics of mind questions it is designed to help us answer. But the good news is that out of the wreckage, a clearer picture of the relevance of different sorts of empirical evidence—and armchair philosophical considerations—emerges.

Chapter 1—Epiphenomenalism and Evolution

I. Introduction

The purpose of the first two chapters of the dissertation will be to present and evaluate an argument that we have overwhelming reason to believe that epiphenomenalism—understood as the doctrine that mental events are ontologically distinct from physical events but causally inert—is strongly disconfirmed vis-à-vis physicalism, based on seemingly uncontroversial facts about the evolution of the human organism.[1]

It is one of the oldest truisms in the philosophy of mind that epiphenomenalism is an extremely counterintuitive and *prima facie* unattractive view, but few attempts have been made to say anything stronger about it than that it is counterintuitive and unattractive. In this chapter, however, I will examine one such attempt. The argument in question tries to show that evident facts about the phenomenological quality of our mental life would be very unlikely if epiphenomenalism were true and much more likely if it were false (and physicalism true). Therefore, these facts about our mental life are thought to serve as strong and perhaps even decisive disconfirmation of epiphenomenalism. I hope to formulate this argument in a more detailed and precise way than it has been previously and evaluate it comprehensively.[2] Ultimately, the discussion

---

[1] For those who think the falsity of physicalism and truth of dualism can be known *a priori*, and so who think there is little point in examining an empirical argument where physicalism is pitted against a dualist option, I invite you to withhold judgment on this issue until the final chapter, when it is discussed. (Or alternatively, the reader can feel free to skip ahead to that discussion now, before beginning.)

[2] It is worth mentioning that traditionally the argument has been presented as an argument against epiphenomenalism *simpliciter*, not an argument against epiphenomenalism when compared with physicalism. There have most likely been a number of reasons for this. One may be that previous philosophers have believed we should settle the dispute between epiphenomenalism and physicalism a

of this argument will propel us to consideration of a broader argument against epiphenomenalism (and in favor of both physicalism and interactionism, not just physicalism), but considering that broad argument right off the bat is unwise. It involves a number of issues that are peripheral to many of the questions we will need to answer, and so it is best to postpone it until the time is right (which will not be until the dissertation's final chapter).

The structure of this chapter will be as follows—first, I will attempt to explicate as clearly as possible what is meant by epiphenomenalism. Second, I will discuss why it is that one might be motivated to adopt epiphenomenalism or to criticize it given that many philosophers, both historical and contemporary, have thought that it is a straw man and not even a minimally plausible doctrine. Third, I will explain what the exact problem for epiphenomenalism is—roughly, that if epiphenomenalism were true the fact that we feel pain under certain conditions and not under others and likewise pleasure under certain conditions and not under others would be utterly mysterious and very improbable. In the next two chapters, which serve as intimately related sequels to this one, I will examine a range of objections to the argument—answering some of them and fine-tuning the argument to deal with others. In the process, I will explore some proposed improvements (and allegedly neutral changes) one might make to the argument, and discuss their impact.

---

priori (if we are able to settle it at all). Another may be that they have thought interactionist dualisms share the relevant advantage over epiphenomenalism with physicalism (namely, causal efficacy of qualia), and so are justifiably lumped together with physicalism for the purposes of the argument. I will discuss these matters in due course—so as not to get ahead of ourselves, for present purposes I will simply adapt discussion of these classic formulations of the evolutionary argument against epiphenomenalism to my particular concern of comparing epiphenomenalism with physicalism. As the discussion progresses, hopefully the justification for this maneuver will be made manifest.

II. What Exactly is Epiphenomenalism?

As I said above, epiphenomenalism as I understand it is basically the doctrine that all our mental event tokens are ontologically distinct from physical event tokens but causally inert, both with respect to other mental event tokens and to physical event tokens.[3] Although this has been the most common understanding of epiphenomenalism historically, in this paper I will focus solely on qualitative mental events, or simply "qualia" or "qualitative events."[4] The reference of the term 'qualitative event' is fairly wide-ranging, however. It is intended to include all phenomenally conscious mental events, including the phenomenological components of occurrent belief and desire.[5] Examples of qualitative events include visual experiences, auditory experiences, pains, and itches.

So, as a first pass at a precise definition of 'epiphenomenalism', we can say that it claims that for any (qualitative) mental event *M* and physical event *P*, *M* is distinct from *P*, and *M* does not cause *P*. (Let the '*M* does not cause *P* ' (for all *P*) be expressed in prose as the *M*'s being 'causally inert with respect to the physical' or 'causally inefficacious with respect to the physical'. Any event that violates this condition will be called 'causally efficacious with respect to the physical'.)

---

[3] I will not deal specifically at the moment with variant versions of epiphenomenalism that countenance mental-mental causation only, such as the one Frank Jackson flirts with in Jackson (1972) and Jackson (1982). These issues will come up as the discussion proceeds.

[4] A note is in order to describe my somewhat unorthodox method of notation in this work. As is standard, I employ single quotation marks when I am mentioning rather than using a term, phrase, or statement (e.g., 'Caesar' has 6 letters). I also employ single quotes for any quotation marks that appear inside other quotation marks. As is becoming more common, I use capital letters when referring to a concept or proposition (e.g., WATER is a vexed concept in philosophy.) The non-standard part of my method is that I reserve double quotation marks for both cited passages and for the standard English scare quote usage. (Context should make the usage apparent.) My reason for preferring the unusual method is precisely that it allows for scare-quoting, which expands expressive power, but is impossible to use in normal philosophical writing since there is no way to convey it without employing some notational device officially designated for another purpose.

[5] I say "phenomenological components" rather than occurrent belief and desire *simpliciter* since, on some views, a person can have occurrent belief or desire without that desire being phenomenally conscious.

The intuitive picture here is one where qualitative events stand to each other and to physical events the way the shadows cast by a moving car stand to the car and its shadows at subsequent times.[6] Although the shadow at time *t* is conjoined in a predictable way, both in terms of shape and position, with the shadow and the car at time *t+1*, the shadow at *t* does not cause either the shadow or the car to have the shape or position that it has at *t+1*. Rather, it is the properties of the car at *t* (or the car instantiating those properties, if you like) that cause both the shadow at *t* and the subsequent car positions and shadows. In an analogous way, the epiphenomenalist alleges that qualitative mental event tokens do not cause other qualitative mental or physical event tokens to occur, although they are conjoined predictably with physical event tokens of specified types (and presumably other mental event tokens as well, though more on this below).[7] So if someone slices my arm with a knife and I subsequently feel tremendous pain and yank my arm away, on the epiphenomenalist view my pain along with the yanking of the arm will just be common effects of the same cause—specifically, the underlying physical brain event that the arm slicing brings about (via intervening events in the nervous system). In no way will the pain have any causal impact on the yanking of the arm or any subsequent mental events associated or conjoined with the yanking of the arm.[8]

---

[6] See Kim (1998) for a fuller discussion of this analogy.

[7] I assume here that epiphenomenalists are willing to claim that mental property types nomologically (or naturally) supervene on physical property types. (Though Frank Jackson may be or at least have been an exception to this at one time, since he took seriously the possibility of mental-mental causation, which is plausibly thought to be in tension with nomological supervenience of the mental on the physical. See Jackson (1982), p. 133). The issue of supervenience is a complicated one, however, and this is not the place to discuss it in the detail it deserves. Consequently, I will let the issue rest for the moment. For more nuanced and detailed discussion, though, see Kim (1998), especially Chapter 1, McLaughlin (1995), Stalnaker (1996), and Sider (1999).

[8] There are, of course, situations where it is fairly clear that pains do not have causal impact on behavior. One example of such a scenario is an unfortunate but perennial favorite among philosophers for discussing mental causation—the case of a hand touching a burning stove. Here, though the subject does feel pain, the

Two obvious difficulties that one faces in making the formal definition of epiphenomenalism fully precise are how to understand the terms 'ontologically distinct' and 'cause'. First, let me deal with ontological distinctness. To clarify the issues, let us make the not overly controversial assumption of realism about properties. That is, let us take it for granted that there really are properties, i.e., that the true comprehensive theory of metaphysics quantifies over them.[9] Now, these properties need not be full blown universals—they could simply be tropes or the like. (We will suppose that they are not merely sets of objects, however.)

The point of these assumptions is to ensure that our metaphysical picture of events is one wherein every event is constituted by an object instantiating a property at a particular time. Two events are ontologically distinct, then, roughly if and only if one event has a different constitutive object, or involves the instantiation, by the constitutive object, of a different property from the other, or occurs at a different time.[10]

There are undoubtedly several caveats, qualifications, and disclaimers that must be added to this definition to make it fully satisfactory. In spite of the difficulties, though, I think it is clear enough intuitively what ontological distinctness amounts to, so I will not belabor the i-dotting and t-crossing unnecessarily. The idea, as it applies to epiphenomenalism, is that there really are mental properties and physical properties that

---

hand is jerked away not as a result of the pain but rather as a reflex reaction via some unconscious mechanism of the nervous system. I intend my discussion in this paper to abstract away from such difficulties, and I have done my best to choose example cases that do not suffer from these relatively trivial but nonetheless potentially confusing problems.

[9] By making these assumptions, I prevent my arguments from coming into dialectical contact with the views of philosophers like Donald Davidson, who struggle with the notion of mental causation but reject the more metaphysically robust ontology that my discussion presumes. See Davidson (1970).

[10] I'd conjecture that talk of properties could be paraphrased into non-property realist terminology while preserving the crux of the difficulty, but I won't speculate further on this here. For present purposes, I'll restrict myself to the less ambitious project of dealing with the property realist.

serve as fundamental constituents of the universe,[11] and that the mental properties (tokens and types equally, if the distinction between tokens and types winds up being a real one) are not identical to the physical properties, nor are they constituted by them, nor do they supervene on them with metaphysical necessity. Moreover, both are instantiated in normal human subjects. On a plausible construal of events, then, this will entail that (e.g.) the event that is my C-fiber firing (where 'C-fiber firing' denotes whatever the physical neural basis of my pain)[12] will be distinct from the event that is my being in pain, since the property (token) of C-fiber stimulation will be distinct from the property (token) of being in pain.[13]

A less technical way to get at the driving intuition here is via the notion of a MINIMAL PHYSICAL DUPLICATE.[14] A physical duplicate of a world *w* is a world that is exactly like *w* in every physical respect. So, for instance, any world is a physical duplicate of the actual world that contains all the physical entities that the actual world does, in all the very same arrangements, with all the same physical laws. (Not just at this moment, of course, but across the entire history of the worlds.) It contains a duplicate of my body sitting at a duplicate of my computer; it contains a duplicate of the Grand Canyon; it contains a duplicate of the Roman Forum on the day Caesar was assassinated, etc. A minimal physical duplicate of a world *w* is a physical duplicate of *w* that contains

---

[11] To say that the mental properties are fundamental is not intended to imply anything about whether or not they are emergent.

[12] Incidentally, whenever I speak of a 'neural base' or 'basis', I intend the states or events spoken of to be understood as purely physical. If there are (non-physical) phenomenal properties had (in some sense) by these neural bases, they will not count as part of the neural bases for our purposes.

[13] I do not want to take on the complicated metaphysical issues surrounding events in any substantive way in this dissertation. I do not think how the details of a plausible theory of events are spelled out will have any noticeable impact on the problem at hand.

[14] As far as I know, minimal physical duplicates are first discussed explicitly in Jackson (1998), though the basic idea is much older. I am grateful to Brian McLaughlin for suggesting a formulation in terms of minimal physical duplicates.

nothing more (by way of objects, properties, laws of nature, etc.) than it would have to in order to be a physical duplicate of *w*.

Now, it is trivially the case that the actual world is a physical duplicate of itself, since after all it is exactly like itself in every physical respect. But it is far from trivially the case that the actual world is a minimal physical duplicate of itself, since it is not obvious that the actual world contains only those properties, events, and laws it would if it were composed only of its physical entities. Many philosophers have suggested that qualia are just the sort of thing that a minimal physical duplicate of the actual world would not contain, in fact. These observations allow us to define both 'physicalism' and 'dualism', and this will be very useful as we go along:

Physicalism := Any world that is a minimal physical duplicate of the actual world is a duplicate *simpliciter* of the actual world.[15]

Dualism := The negation of physicalism

Now that we have examined ontological distinctness and related ideas surrounding minimal physical duplicates, we are halfway to being able to formulate a workable and reasonably precise definition of epiphenomenalism. Only an analysis of causation stands in our way. Causation is, of course, one of the most vexed and

[15] As a fully satisfactory definition of what is typically called 'physicalism', this formulation might run into some difficulties. The most serious one appears to be that the definition doesn't state an intuitively sufficient condition for physicalism, since it is compatible with the existence of necessarily existing non-physical entities, such as God. In any case, though, it does state a significant necessary condition for physicalism, and one which all philosophers who would intuitively qualify as dualists in our sense would deny. Hence, since capturing the relevant distinction in the mind-body sphere is what we are concerned with (not with articulating a precise understanding of physicalism as a thesis about the whole of reality), the definition should work fine for our purposes. (I am grateful to Brian McLaughlin for pointing out the need for this qualification.)

complicated of metaphysical topics, and I do not intend to enter the fray here in any substantive way. I will merely aim to give an intuitive characterization of what causation amounts to in a way that captures what most epiphenomenalists (and realists about mental causation for that matter) think it amounts to.[16] The notion of causation that typically gets assumed in the debate over mental causation is a very common sensical one—one where causation requires the production or generation of, or (for lack of a better word) direct "oomph" on the effect. According to this picture, instances of causation will not be reducible to facts about counterfactual scenarios, though they may entail such facts. At the very most, if causation is not a fundamental feature of the world, instances of it will be correctly explained by robust causal laws which will themselves be fundamental features of the world. So, *ceteris paribus* epiphenomenalism will be true if and only if causal relations do not obtain between mental properties or events and other properties or events (with the mental properties/events in question as the causes).

It is natural to wonder why I am restricting my attention to those who hold a general theory of causation which is robust in this fashion, since other views are certainly held. There are two main reasons. The first is sociological; as I said, it is fairly common in debates about mental causation—much more so than in, say, general metaphysical debates about causation—for all the parties to maintain a shared background conception of causation. Typically, this shared conception of causation is a robust, "oomphy" one. Second, it is much harder to make sense of the distinction between events that are epiphenomenal and those that are not on other understandings of causation.[17]

---

[16] Consequently, I won't deal with those who object to epiphenomenalism on the grounds that the notion of robust causation *simpliciter* is problematic, aside from a few passing remarks. For a view like this, see Loewer (2001).

[17] Broad (1925) makes a similar point.

Consequently, it is convenient to assume a theory of causation that is friendly to drawing the distinctions that we will be highlighting. If other theories are able to recover intuitive verdicts about epiphenomenal and non-epiphenomenal events, then I suspect the results of our discussion will apply equally to them *mutatis mutandis*.

Now that we have gone through at least a basic discussion of the potentially tricky concepts surrounding epiphenomenalism, we are in a position to offer a succinct and reasonably precise definition. It goes as follows:

Epiphenomenalism := Dualism and all qualitative events are causally inert with respect to the physical

III. Why Would One Want To Be An Epiphenomenalist in the First Place?

Now that we have set the preliminary groundwork and achieved a more precise understanding of the view, my first task main task will be to explain why one might be motivated to attack epiphenomenalism or be an epiphenomenalist given that it is regarded by many as a straw man and an immensely implausible one at that.

It is no secret that the philosophical literature is replete with opposition to epiphenomenalism on account of its counterintuitiveness and offensiveness to common sense. Richard Taylor, for instance, calls epiphenomenalism "so bizarre a description of human nature as to make almost any alternative conception more acceptable."[18] No less than Jerry Fodor has said that "if it isn't literally true that my wanting is causally responsible for my reaching… and my believing isn't causally responsible for my

[18] Taylor (1992), p. 26.

saying… then practically everything I believe about anything is false and it's the end of the world."[19]

Implausible though it may be, however, epiphenomenalism is not a straw man if what is meant by 'straw man' is a view that no one holds. Historically, figures such as Malebranche, Huxley,[20] and perhaps Leibniz endorsed it, and in recent philosophy it has been embraced by Frank Jackson[21] and William Robinson.[22] In addition, it has been taken very seriously by a number of others, most prominently David Chalmers.[23] Interest in epiphenomenalism has grown in recent years in fact, largely because of Chalmers' ground-breaking work on property dualism and his consequent flirtations with the epiphenomenalist view.

Philosophers are normally driven to epiphenomenalism (in spite of its universally recognized counterintuitiveness) because of two broad metaphysical doctrines that are, taken individually, at least fairly plausible. When someone holds both of the views with more conviction than she holds the intuition that the mental realm has causal efficacy, the epiphenomenalist position becomes attractive.

The first of the views is dualism, which have already discussed. Although dualism has a history as old as philosophy itself, recent decades have seen the proliferation of pro-dualist arguments, led by Saul Kripke[24], Jackson[25], and Chalmers[26].[27]

---

[19] Fodor (1990), p. 156. Although this particular reference is to mental entities that are plausibly construed as non-qualitative, it is not hard to feel its force when extended to qualitative ones.
[20] See Huxley (1874)
[21] In Jackson (1982)
[22] See Robinson (2004b)
[23] See Chalmers (1996), especially Chapters 4 and 5.
[24] Kripke (1972), especially Lecture III.
[25] Jackson (1982) and(1986).
[26] Chalmers (1996), though it is not altogether clear that Chalmers counts as a dualist on this understanding of dualism, since he considers the possibility that mental properties may be the intrinsic stuff of physical properties that are ordinarily specified only relationally.

I will not recount these familiar arguments here, but suffice it to say that interest in the view has grown after a prolonged lapse.

The second of the doctrines is causal closure of the physical or, as it is sometimes put, causal "exclusion" of the physical.[28] This is the view roughly that every physical event has as its actual sufficient cause another physical event (with a caveat here and there to allow for cases of indeterministic causation, such as perhaps that every physical event which is made more (or less) probable is made more (or less) probable only by physical events).[29] As Jaegwon Kim puts it, "[i]f you pick out any physical event and trace out its causal ancestry or posterity, that will never take you outside the physical domain."[30] Incidentally, the popularity of the causal closure view is a long-standing one, particularly in the neurosciences. In 1870, for instance, in his lecture to the Imperial Academy of Sciences in Vienna, Ewald Herring stated that brain physiologists should make "the unbroken causative continuity of all material processes an axiom of [their] system of investigation."[31]

Reasons for holding the causal closure view are not often articulated very explicitly, but typically they have to do with faith in the in-principle completability of physics—that is, the belief that there can be a "complete and comprehensive theory of all physical phenomena,"[32] where this comprehensive theory presumably captures what is truly occurring and does not posit entities outside the physical domain. (Toward the end

---

[27] Though many of these arguments have been in the air in the relatively recent history of philosophy (albeit in less sophisticated forms). See, for instance, Alexander (1920), Broad (1925), and Russell (1927). Also, for an alternative version of similar pro-dualist arguments, see Rosenberg (2004).
[28] For a defense, see Kim (1998) and Chalmers (1996).
[29] Thanks to John Hawthorne for pointing out the need for such a qualification and to Dean Zimmerman for suggesting the content of the qualification.
[30] Kim (1998), p. 40.
[31] See McLaughlin (2006)
[32] Kim (1998), p. 40.

of the work, we will have occasion to examine causal closure and some related theses in more detail.)

Another common motivation for accepting the causal closure view is a refusal to countenance causal overdetermination (roughly the idea that there can be more than one event that is the actual immediate sufficient cause of another event).[33] [34] Although many dualists who believe in causal closure of the physical leave their adherence to this motivating doctrine implicit (and leave attribution of it to them as the work of charitable interpreters), Chalmers does an admirable job of bringing his (at least tentative) commitment to it out into the open and arguing for it.[35]

We are now in a position to see more clearly why epiphenomenalism has been seen as a live option in spite of its counterintuitiveness. Dualism and causal closure together entail that qualitative properties cannot be causally efficacious with respect to physical properties, and by extension that qualitative events cannot be causally efficacious with respect to physical events. The reason is that, by dualism, qualitative properties (and events by extension) will be ontologically distinct from physical

---

[33] "Overdetermination" in this context is to be understood as robust overdetermination—overdetermination by properties that are ontologically distinct from one another, in the sense discussed earlier. (A somewhat more detailed exploration of the problems posed by overdetermination will accompany the examination of interactionism later in the work.) Overdetermination worries brought on by supervenient properties that are metaphysically necessitated by their supervenience bases are not relevant. For our purposes, we can feel free to claim that they, and the causal relations they enter into, are an "ontological free lunch," in the words of Chalmers (1996), p. 179. For a discussion of them, though, see Yablo (1992).

[34] Causal overdetermination is made much more palatable when one accepts non-robust conceptions of causation, including especially Humeanism—roughly the claim that As cause Bs iff they are regularly conjoined with Bs and precede them temporally. Hence, part of the explanation for my desire to avoid considering non-robust theories of causation as much as possible.

Incidentally, I will not attempt to deal with the potential implications of mental causation problems on sophisticated Humeans like David Lewis who do not believe in robust causal relations but who nevertheless refuse to acknowledge that just any constantly conjoined events with the right sorts of temporal relationships stand in a causal relation. See, for instance, Lewis (1986). As I said above, insofar as these philosophers can preserve the intuitive distinctions between epiphenomenal and non-epiphenomenal events, my conclusions should apply to them as well, but I won't attempt to demonstrate that here.

[35] As does Kim, though he may not be a dualist. For Chalmers's arguments, see Chalmers (1996), Chapters 2 and 4, especially pp. 74-75, 86, and 151-156.

properties, and by causal closure, all physical events will have only physical events as their (sufficient) causes.

Though the doctrines together do not entail the impossibility of qualitative events having causal impact on other qualitative events, they are in significant tension with the view that qualitative events do have causal impact on other qualitative events. The reason is parallel to the one in the mental-physical case. If physical events (in the brain) or physical properties (of the brain) are causally responsible for qualitative events or the instantiation of qualitative properties (as might plausibly be thought if dualism is true, and as virtually all dualists acknowledge), then there will be no causal role for other qualitative events to play in the process. The problem is exacerbated—or at least made more stark and clear—when holders of the doctrines endorse the view that there are true conditionals of the form "if physical event of type *p* then mental event of type *m*" (where *p* and *m* type events by their natural qualitative properties) with a modality stronger than a material conditional. In other words, when holders of the doctrines acknowledge (as philosophers like Chalmers do) that the mental supervenes on the physical because of some lawful or causal connection between the two (and not merely because they fortuitously happen to coincide), it becomes very difficult or even impossible to coherently resist the conclusion that mental events have no causal efficacy on other mental events. This is because it looks like physical events in the brain are doing all the causing of mental events, leaving no room for mental events to do any causing of their own.

Because the purpose of this chapter is primarily to discuss an argument against epiphenomenalism and not to discuss the motivations for being attracted to

epiphenomenalism in the first place, I will not attempt to pursue the issue in any further detail, though there remains much to be said.[36] In this section, I have simply tried to show in outline why epiphenomenalism might be thought plausible in spite of its counterintuitiveness and why the project of criticizing it is worth pursuing. In the next section, I will delve into the specific argument against it.

IV. The Problem for Epiphenomenalism

The argument I will examine against epiphenomenalism is based on uncontroversial background assumptions about human evolution—namely, that human beings evolved and that natural selection was a significant factor in that evolution, in particular the evolution of the brain and mental life.

Traditionally, most epiphenomenalists and those who take epiphenomenalism seriously have not believed that the fact that human beings evolved has any evidential bearing on the epiphenomenalism question. Jackson, for instance, considers the possibility that because consciousness evolved we thereby have evidence that it is causally efficacious, since only if a thing is causally efficacious can it make a difference to behavior and thereby improve its chances of being selected for. In the end, though, he concludes that consciousness is probably just an inevitable by-product of "certain brain processes that are highly conducive to survival."[37] Chalmers, in *The Conscious Mind*, also considers the evolution question, but provides a similarly dismissive response. He says that "like the fundamental laws of physics, psychophysical laws are eternal… It may be that in the early stages of the universe there was nothing that satisfied the physical

[36] For excellent and much fuller discussions of the crux of the tension, see Chalmers (1996), pp. 161-168, and Kim (1998), pp. 38-47.
[37] Jackson (1982), p. 134.

antecedents of the laws, and so no consciousness… [but] when [physical systems that satisfied the relevant conditions] came into existence, conscious experience automatically accompanied them by virtue of the laws in question…"[38]

Some philosophers who take epiphenomenalism to be a serious dialectical option, however, have believed that these and similar responses, orthodox though they are, fail to do the evidential significance of evolutionary considerations justice. This group has included Herbert Spencer, William James, and Karl Popper.[39] Despite the impressiveness of the names who have weighed in on the issue, their formulations of the problem and the argument spawned from it often leave much to be desired in the way of precision, and require quite a bit of charitable interpretation.

Consequently, I will devote the remainder of this section to explaining in detail why these philosophers have believed dismissive responses like Jackson and Chalmers' above to be mistaken, hopefully filling in the gaps in their accounts and formulating their claims in a more plausible and precise way.

---

[38] Chalmers (1996), p. 171.

[39] See Spencer (1870), (1871), (1883), James (1879), (1890), and Eccles and Popper (1977). For a response, see Broad (1925). The argument was also formulated very early on, in 1882, by G.J. Romanes. See Romanes (1896) for a restating.

Let us begin by imagining a case where someone cuts me in the arm with a knife.[40] Also, suppose epiphenomenalism is true. In this scenario, the very intense pain I feel when the knife cuts me in the arm will have no causal impact on my subsequent behavior—most likely a jerking of the arm away. (The pain property token will have neither a direct causal impact on that behavior nor an indirect one in terms of causing any phenomenal judgments or intentions.)

Before proceeding with any further argument, let me draw some background distinctions and make some supporting claims. First, the proponent of the evolutionary argument will claim that this token of pain is associated with an intrinsically negative phenomenology. By 'intrinsically negative phenomenology' (hereafter INP),[41] I mean here something stronger than that as a matter of psychological fact any normal subject in the "acquaintance" relation (or "having" relation) to the token or a qualitatively identical one is *ceteris paribus* disposed to be averse to its environmental or bodily cause (i.e., desire its absence).[42] (The "normal" qualification is intended to rule out counterexamples involving persons with masochistic tendencies, for instance, and the "*ceteris paribus*" one to cover cases where persons do desire the phenomenology or its cause for instrumental reasons, in virtue of some desired state of affairs that is brought about or made possible as a result.) Although a subject having an INP will as a matter of fact typically always be averse to it in this way, the proponent of the evolutionary arguments (for reasons that will be discussed later) will not acknowledge that the actual having of

---

[40] This example is descended from one proposed by Chalmers in conversation.

[41] Please consult the appendix at the end of the dissertation for a glossary of major acronyms. The glossary contains each commonly used acronym, the chapter of its introduction, and an informal definition.

[42] For clarity's sake, here I assume what is often called the "quality-based" view of experience—where experiences consist of the instantiation or having of a phenomenal property by a subject. There are other possibilities—such as a sense-datum view or a Humean view (where there are no subjects, only "free-floating" experiences). For our purposes, the choice is irrelevant, and any reader who prefers a different view can feel free to paraphrase the argument into the categories appropriate to that theory.

the INP is analyzable in these terms. Rather, the having of the INP will be something much more fundamental and primitive. It is difficult to say anything illuminating about INPs, but as a first take, it's probably fair to claim that to have an INP is simply to be in a state that feels *bad*, where the badness is part of the phenomenology itself.[43] (Later, I will discuss whether anything more illuminating than this about the intrinsic negativity can be said.)

There is of course more to be said generally about the subject's having of this pain token and the phenomenology that constitutes it—for instance, the token is a member of a wide type: the type under which all tokens fall such that normal persons, when having them or acquainted with them, are disposed to be averse to them *ceteris paribus*, and find them negative. These include itches, aches, pains, depression, anxiety, hunger pangs, sickness phenomenologies, negative drug phenomenologies, etc. The token also belongs to a narrower type—the type that encompasses what would intuitively be called "pains" *simpliciter* (i.e., bodily pains). I am not sure what exactly distinguishes pain from other types of mental events associated with INPs, nor am I sure the distinction between a pain and an unpleasant itch, for instance, is not extremely vague. There does seem to be a clear intuitive distinction (albeit vague perhaps) between pains and other types of mental events associated with INPs, however. (This distinction may be a primitive introspective one, not subject to any illuminating analysis.) The token also belongs to a narrower type still—the type that might be characterized as a "sharp pain" as opposed to a "dull" one. These metaphorical descriptions are at best useful heuristics for

[43] I use the term 'bad' with hesitation, because I do not wish to suggest that there is any straightforward connection between this sort of phenomenology and moral badness. Unfortunately, all the typical English words I could appropriate for this use are in similar danger of being construed as having moral implications. Short of inventing a pure term of art for the purpose, there is no alternative but to acknowledge and remain aware of the potential for confusion on this point.

indicating which fine-grained phenomenal pain concepts we are gesturing at, but intuitively I think we all have a grasp of the distinctions between various sorts of pain phenomenologies. For the time being, we will concentrate only on the general character of the phenomenology—intrinsically positive or intrinsically negative. Later, we will investigate the relevance of considering more fine-grained evidence than just the general positivity or negativity of the qualia, such as (for instance) its being a sharp pain of a particular type.

To return to the argument, assuming epiphenomenalism is true, the painful sensation I feel when my arm is cut has no causal impact on my subsequent behavior. However, assuming also that I am the product of an evolutionary process, it seems sensible to ask "why are tokens of this type (i.e., intrinsically negative) correlated so nicely with events like the arm cutting, events that are uncontroversially harmful to my prospects for survival?"[44] The common sensical explanation (clearly available to the physicalist) is that these qualitative properties have a direct causal impact on my current and subsequent avoidance behavior, or at least have an indirect causal impact (perhaps probabilistic) on it via an alteration of my conscious judgments and intentions. Consequently, the ones that dispose me to engage in the avoidance behaviors are likely to be selected for, since they are the ones that make it more likely that I will survive whatever travails confront me. But at this point, the proponent of the evolutionary argument will claim that this explanation is not available to the epiphenomenalist, since

[44] Most classical versions of the argument, especially James's, have relied on correlations between broad phenomenological type and behavior/distal stimulus. (E.g., INP with cut to the arm/jerking away of the arm). Consequently, it is those I will focus on for the time being. Later, I will distinguish more carefully between arguments that employ broad phenomenological evidence and ones that employ more fine-grained phenomenological evidence (and other kinds of fine-grained evidence as well). The arguments I ultimately defend will be ones that employ such fine-grained evidence, but I do not consider them until after discussion of arguments modeled on the historical ones is complete.

the epiphenomenalist denies that qualitative events play this sort of causal role in behavior. The defender of the evolutionary argument will then allege that it is a mystery why these mental events associated with INPs are correlated with the physical events that they are.[45]

The defender of the argument might urge that there is more evidence to take into account as well. In general, it seems the more survival-threatening the stimulus event, the stronger the INP associated with it. Typically, deep wounds are much more painful than tiny scratches *ceteris paribus*, for instance. (Exceptions, though, include things like exposure to radiation, which historically were not part of the evolutionary environment.)[46] Maybe the epiphenomenalist could plausibly urge that this evidence isn't such a big deal, though, since there is probably more pronounced and dramatic activation of the relevant neural structures, and this magnification could be reflected in the bridge laws, yielding a stronger phenomenology.[47] Nevertheless, it's something to keep in mind.

It is a very fundamental tenet of confirmation theory that a piece of evidence confirms a hypothesis h_1 relative to another hypothesis h_2 (i.e., makes h_1 more likely to be true relative to h_2) if and only if the ratio of the probability of h_1 over the probability of h_2 is greater on the evidence (plus background assumptions) than is the ratio on the lack of the evidence. In turn, this is true if and only if the evidence is more likely on h_1 than on

---

[45] The defender of the argument will allege that this problem is only made worse for the epiphenomenalist when we notice that mental states associated with INPs can also be relevantly typed according to their position on the body. Why, we might wonder, does a burn on the leg produce the same sort of phenomenology as a burn on the arm? Perhaps the answer has to do with the wiring of our nervous system, but these considerations are just one more worry for the epiphenomenalist. (I am indebted to Colin McGinn for this point.)

[46] except from the sun, though this is nitpicking

[47] Chalmers, for instance, speculates that the intensity of experiences is dependent upon the extent to which the underlying neural state plays a "control role" in the overall activity of the brain. See Chalmers (1996), p. 224.

h_2. (In other words, the ratio of the probability of the evidence on h_1 over the probability of the evidence on h_2 is greater than 1.)[48] To put it more formally:

$$\frac{P(h_1/e \text{ and } k)}{P(h_2/e \text{ and } k)} > \frac{P(h_1/k)}{P(h_2/k)} \quad \text{iff} \quad \frac{P(e \text{ and } k/h_1)}{P(e \text{ and } k/h_2)} > 1$$

Here e is the evidence, k is the background knowledge, and h_1 and h_2 are hypotheses.[49] The inequality on the left is intended as an account or analysis of what it is for h_1 to be confirmed relative to h_2.

Applying this result to our current case, the proponent of the evolutionary argument will claim that epiphenomenalism is disconfirmed relative to physicalism if and only if the evidence that we actually observe is less likely if epiphenomenalism is true than if physicalism is. The epiphenomenalist contends that the evidence (i.e., the correlations between phenomenological type and distal stimulus) is not less likely on epiphenomenalism, but the defender of the argument has offered what he believes are convincing reasons to think otherwise, and consequently convincing reasons to think epiphenomenalism is disconfirmed. To repeat: these reasons are that the observed correlations between qualitative mental events and events like cuttings of the arm are supposed to be very likely on physicalism, since physicalism allows for mental causation of behavior, but unlikely on epiphenomenalism. If physicalism were true, we would be

[48] For a comprehensive introduction to the subject of confirmation, see Swinburne (1973) and Howson and Urbach (1996). There are, of course, numerous qualifications to this thesis, but none of them are relevant for present purposes.

[49] Probability and likeliness here are to be understood as subjective or epistemic probability, not something like an objective chance. I will not attempt to tackle these very complicated probability issues here, however—I think the relevant notions are clear enough intuitively for the limited current purposes.

led to expect roughly the correlations we find, whereas supposedly if epiphenomenalism were true there would be no reason to expect these correlations.[50] Rather, if epiphenomenalism were true, there would be no reason to think that horribly survival-threatening events wouldn't be associated with the most sublime of pleasures. This is because if epiphenomenalism were true, then phenomenology would have no causal impact on behavior, and behavior is what natural selection selects for. (At this point, the reader may be wondering about the likelihood of the evidence on interactionist versions of dualism as well, a piece of the puzzle required to come to a decision on the *overall* confirmation of the various hypotheses. As promised, this discussion will be coming later on in the work. For now, though, we will be restricting our attention to only physicalism and epiphenomenalism. Later, other possibilities will be explored.)

[50] It should be noted that, given our definition of 'physicalism', both role and filler functionalist views count as physicalist theories.

Chapter 2—Does The Evolutionary Argument Work?

Now we have seen the evolutionary case against epiphenomenalism laid out in more detail, and hopefully more precision, than it has been laid out historically. As we have seen, the argument claims that human evolution provides a major stumbling block to would be epiphenomenalists, because the correlations we observe between survival-threatening events and phenomenology would be very improbable if epiphenomenalism were true, but likely if physicalism were. Consequently, we allegedly have good and perhaps conclusive reason to prefer physicalism.

Naturally, the issue we must now face is whether the evolutionary argument has the anti-epiphenomenalist dialectical force that its defenders have believed it does. As I stated above, I will ultimately argue that it does not. Our efforts will not be wasted, though. The process of examining both the narrower evolutionary argument formulated in the previous chapter (the one pitting epiphenomenalism against only physicalism) and broader ones later in the work will propel us toward a greater understanding of the interplay between empirical and conceptual considerations in debate over the mind-body problem.

Over the course of evaluating the narrower anti-epiphenomenalism argument, I will examine a number of objections, roughly in order of plausibility from least to greatest. (Some of the objections challenge the "big picture" of the argument and its underlying strategy, while others just challenge the details.) I will argue that the proponent of the argument can readily provide answers to most of the earlier objections, but that some of the later ones are potentially problematic. I don't believe any of the objections before the final one constitute a devastating challenge to the overall strategy,

though. When the going gets tough with some of the other objections, I will sketch ways that the details of the argument can be changed to cope with the difficulties posed. (We will ultimately see, in fact, that the objections challenging INPs and IPPs in various ways are ultimately irrelevant, since there are amended versions of the argument that don't employ evidence that appeals to qualitative events belonging to broad phenomenological categories like INP and IPP.)

To preview coming attractions, after examining (all but one of) the various potential objections to the argument in the remainder of this chapter, in the next chapter I will examine alterations that can be made in the general form of the argument, and explore whether they help or hurt the prospects of the overall argument strategy. Later, I will introduce interactionist forms of dualism into the mix (as historical formulations of the evolutionary argument typically have), and discuss the as yet unexplored issue of whether they are vulnerable to similar arguments as epiphenomenalism (rather than benefiting from them, as has often been thought by defenders of the argument form). At that point, I will also consider the final objection to the general argument strategy that I alluded to above. This objection cannot be fully appreciated until all the options are on the table, but once they are, we are able to see how the objection ultimately dooms the argument.

Without further ado, let us proceed on to a discussion of the objections. The ones that I will examine are as follows (the very rough descriptions below will be replaced by thorough explications as we proceed through the discussion of each objection in turn):

(A) *The argument trades on conceptual rather than empirical considerations.*

(B) *There are survival-threatening stimuli associated with IPPs and stimuli that enhance the prospects of survival associated with INPs. This undermines the claim that the correlations really are smooth between (e.g.) INP and survival-threatening stimuli.*

(C) *There is good (introspective + scientific) reason to suppose that any phenomenology that would supervene upon the actual neural bases of INPs or IPPs would be negative or positive respectively according to any of the mind-body theories.*

(D) *We really have no concept of an INP or IPP, or at least no introspective reason to suppose any actual qualitative event falls under either of these concepts or any of their more determinate species. Rather, the concepts we might naively associate with INP (like SHARP PAIN, for example) or IPP (like INTELLECTUAL PLEASURE) work in some other way.*

(E) *Although we may have concepts of INPs and IPPs and introspective reason to suppose that actual qualitative events fall under them, scientific findings about the neural bases of what we might naively describe as IPPs and INPs show that nothing really does fall under them.*

(F) *What entitles us to use other people's (both current and historical) phenomenology-stimulus correlations as evidence? Nothing. After all, even if other people had correlations totally opposite from what we normally suppose (e.g., great pleasure with severe burns), if epiphenomenalism were true then, according to the argument's own principles, these people would behave in exactly the same way as they do.*

(G) *If epiphenomenalism were true, how would we know our introspective judgments (and the judgments about the correlations between phenomenological type and distal stimuli) were correct? We would not, so we can't use them as evidence.*

(H) *How would we know there was the right sort of unbreakable connection between intrinsic qualia positivity/negativity and behavior on physicalist theories? This connection is clearly lacking on epiphenomenalism, but why suppose it is any less problematic on physicalism (or any other view for that matter)? There is no reason to suppose so.*

As promised, let us proceed through each of these in turn. I will argue that objections (A)-(C) and (E)-(F) can be readily answered, and highly plausible responses to (D) can be outlined, though I cannot (in a work of this size) complete a full rebuttal. (Even so, the discussion of (D) will be quite lengthy.) (G) and (H), on the other hand, are more difficult objections. While I cannot provide a full reply to (G), I will suggest some avenues of response that hopefully make it clear that (G) is far from a decisive objection to the general argument strategy, and indeed one that must be developed at considerable length before it can have any promise. (H) is a deep worry, and I will ultimately argue that it vitiates the force of the argument we have considered, as well as many variations on the theme of that argument. But far from being a depressing finding, this objection provides the key to appreciating the interplay between conceptual and empirical factors in debate over the mind-body problem. (Much of the discussion of (H) will have to wait till

later chapters, as I said. This is precisely because of its depth, and the corresponding need for more tools to be developed to examine it adequately.)

I. The Preliminary Objections

We will begin with (A). In its full detail, (A) claims that the evolutionary argument does not really rely on empirical considerations to attack epiphenomenalism, but rather adduces conceptual reasons to prefer physicalism.[51] Specifically, it is alleged that the argument's primary qualm with epiphenomenalism is just that it would introduce intolerable complexity in the number and variety of bridge laws between physical/functional and qualitative, and that we ought to prefer simpler hypotheses to more complicated ones *ceteris paribus*.[52]

We can easily see that this objection rests on a misunderstanding of the logic of confirmation. I suspect what defenders of this objection have had in mind is that there is a very specific version of epiphenomenalism that is clearly not disconfirmed by the empirical evidence—namely, the one that posits fundamental bridge laws yielding exactly the correlations between phenomenology and distal stimuli that are actually observed. It is true that the evolutionary argument's only objection to this specific form of epiphenomenalism is that it is incredibly complex and *ad hoc* (particularly if the neural correlates of the various phenomenologies are fairly diverse). Hence, (on normal ways of setting prior probabilities that privilege hypotheses that are either intrinsically simple or cohere well with our background knowledge about the world, or both) this

[51] This is different from the plausible worry discussed in footnotes very early on—that the issue between epiphenomenalism and physicalism (or at least between dualism and physicalism) must be settled *a priori* if settled at all. That issue will be taken up later. The present objection is a less deep, more humdrum claim about the way confirmation principles are being employed in the argument.

[52] This objection was first raised by an anonymous reviewer of an earlier draft of this material from *Pacific Philosophical Quarterly*.

gerrymandered version of epiphenomenalism will come out as extremely unlikely both before and after we take the evidence into account. However, it does not follow from this point alone that epiphenomenalism in general is not disconfirmed by the empirical evidence.

What I refer to as 'epiphenomenalism' is a disjunction of all the various specific epiphenomenalism hypotheses, just as what I refer to as 'interactionism' and 'physicalism' are disjunctions of all the various specific interactionist and physicalist hypotheses respectively. According to the evolutionary argument, epiphenomenalism is severely disconfirmed, since many of the specific epiphenomenalism hypotheses (including virtually all of the ones with relatively high prior probabilities on any reasonable setting of those probabilities) are disconfirmed by the evidence. This is because they supposedly lead us to expect correlations between phenomenology and distal stimuli that we do not in fact find.

For a more intuitive pass at just what is wrong with this objection, consider the following case, which I believe is exactly analogous to the epiphenomenalism one in all the relevant respects: Suppose I know Tom doesn't work at a bank and that he has no ordinary, run-of-the-mill reason to be near bank safes. Now, imagine a bank safe gets robbed and I learn Tom's fingerprints are on the safe. Suppose I consider the hypothesis "Tom didn't rob the bank safe." Is this theory disconfirmed by the fact that his fingerprints are on the safe? Obviously—since Tom doesn't work at a bank and has no ordinary reason to be near a bank safe, if he didn't rob the safe, we wouldn't expect his prints to be on it.

Imagine now an objector claiming something like the following: "really, your argument isn't an empirical one that Tom didn't rob the safe. Rather, it trades on conceptual considerations, aiming to show that a hypothesis where Tom didn't rob the safe is extremely complex. This is because there is a specific version of the "Tom didn't rob the bank safe" hypothesis that is not disconfirmed by the finding of his prints on the safe. This hypothesis might be something like—'the real robber snuck into Tom's house and unbeknownst to Tom made a mold of his hand and rigged up a contraption to leave replicas of his fingerprints on the safe, so as to mislead the authorities.'"

True, this specific hypothesis is not disconfirmed by the evidence, but because it is only one among many "Tom didn't rob…" hypotheses, its ability to survive the evidence without disconfirmation doesn't have much effect on the fate of the general hypothesis. It is the same with epiphenomenalism—just because one very specific (and very gerrymandered) epiphenomenalist hypothesis is not disconfirmed does not imply much of anything for epiphenomenalism generally.

So much for objection (A); let us move on to objection (B). Recall that a critical part of the original argument is that there be very smooth, reliable correlations between INPs and survival-threatening distal stimuli (and the same *mutatis mutandis* for IPPs and stimuli that are helpful for the prospects of continued survival). This evidence is crucial because the central claim of the argument is that epiphenomenalism does not lead us to expect the smooth correlations between phenomenology and distal stimulus (because phenomenology does not cause behavior according to epiphenomenalism), while physicalism does.

But, as the defender of (B) would point out, plainly there are survival-threatening events that are associated with IPPs, in particular bodily pleasure. These include smoking, drug use, ingesting antifreeze or dissolved lead, lying out in the sun for moderate periods of time, living an inactive lifestyle (at least in the short run), eating fatty or sugary foods, etc.[53] Recent evidence even suggests that the beloved "new car smell" may be dangerous, with the chemicals responsible for the scent having noticeable carcinogenic effects even in relatively small doses. Also there are survival-conducive events that in some circumstances, at least, are associated with INPs, such as vigorous exercise, taking certain medicines, and eating healthy food.[54]

These evident facts could be used to form an argument that the probability of the observed correlations are not particularly high on any of the competing mind-body theories either, and so we should not be worried that they turn out very improbable on epiphenomenalism.[55] In fact, one could even imagine using the evidence to construct an evolutionary argument in favor of epiphenomenalism, rather than against it! The idea would be that the correlations between phenomenology and distal stimuli are quite mixed when we look closely—sometimes INPs are associated with noxious stimuli, other times with beneficial ones, and the same for IPPs. If these correlations were indeed very mixed, the defender of epiphenomenalism could contend (for the reasons adduced above

---

[53] Other kinds of activities that are associated with IPPs but are harmful to survival prospects, such as promiscuous sexual behavior, may be explained by the tendency of the behavior to increase reproductive success, even if decreasing the chances of individual survival. Generally the two goals coincide, but there are certainly exceptions. Rather than dealing with the exceptions here, I will simply set them aside. This should not have any noticeable effect on the overall argument.

[54] We are restricting our attention here primarily to simple experiences like somatic feels and very brute taste phenomenology. Undoubtedly the character of some more sophisticated experiences are heavily influenced by cultural factors (which all of the mind-body theories would presumably have similar accounts of), which makes them less relevant for the purposes of the argument we are currently considering.

[55] William James considers an objection like this in James (1890). His example is the pleasure of drunkenness.

as part of the evolutionary argument) that in fact they would be exactly what we would expect to find if epiphenomenalism were true, but not if physicalism were. Consequently, they would support epiphenomenalism (by the same principles employed in the original argument)! But even if the correlations were only somewhat mixed, a defender of epiphenomenalism might at least want to claim that they would not count very heavily in favor of alternative hypotheses—maybe not in favor of them at all.

I would reply, though, by pointing out that these event types are the rare exception rather than the rule. Evolution cannot be expected to adapt us perfectly to the survival challenges of our current environment, both because that environment differs from the ones humans evolved in historically and because there are factors at work in evolution other than pure natural selection (such as genetic drift, for instance).[56] If proponents of the evolutionary argument are correct in their basic assumptions, physicalism, although it doesn't account for every observed correlation, leads us to expect the vast majority of them. Epiphenomenalism, on the other hand, does not, as we have seen. Consequently, it fares far worse than the alternative once we grant that the general strategy of the argument is sound.

Objection (C) contends that a different fundamental assumption in the original argument is mistaken. The evolutionary argument relies on the implicit premise that if epiphenomenalism were true, pretty much any phenomenology could be connected to the actual neural basis of a phenomenological type. The fundamental laws of nature that

[56] I won't attempt to address the possibility that our cognitive mechanisms may be risk averse in a way that provides us with "false positives" in some situations—i.e., a harmless or even helpful stimulus causes an INP in us, because it was to our evolutionary advantage (because of efficiency, perhaps, or close similarity to dangerous stimuli) to have systems that sometimes dispose us to avoid helpful stimuli, because the reward of taking advantage of the stimulus is outweighed by the risk of being harmed by a superficially similar dangerous stimulus. (Among psychologists, this phenomenon is known as the "Garcia Effect.") For a discussion of related issues, see Stich (1983).

govern phenomenology's causation by or emergence from its neural basis would simply have to be different, and most views about the laws of nature see them as contingent.[57] The reason the argument relies on this premise is that if there were some recognizable metaphysical reason why (e.g.) the neural basis of bee sting phenomenology was specially suited for negative rather than positive bridge laws from physical/functional to qualitative, the range of epistemic possibility space (i.e., the space of confirmation) would have to reflect that information. Specifically, epistemic possibility space would have a greater region (potentially a much greater region) devoted to those epiphenomenalist possibilities where that particular neural basis was connected with an INP.[58] This would result in versions of epiphenomenalism with that particular correlation between neural basis and INP receiving a higher prior probability than others, and hence versions of the view with the particular correlation between INP and distal stimulus would as well. The reason is that the connection between distal stimulus and underlying neural basis of the phenomenology, being a completely physical process well accounted for by science, is taken for granted. (It is taken for granted not only by epiphenomenalism but also by the competing alternatives—including physicalism—since none of them posit a causal role for phenomenology between the event that is the organism receiving the distal stimulus and the event that is the tokening of the neural basis of the phenomenology.)

---

[57] The contingency of the laws of nature is really irrelevant here, though. Even on views like Sydney Shoemaker's, where the laws of nature are necessary, different properties could have been instantiated, which would have interacted with the laws in slightly different ways to produce the desired effects. (See Shomemaker (1980).) We would only encounter a problem if both the laws of nature and the properties instantiated were necessary, and very few people take such a view seriously. For a good general discussion about the pros and cons of Shoemakerian views about laws, see Hawthorne (2001).

[58] Later, I will introduce much more elaborate metaphors for thinking about epistemic possibility space, but what I have said here should suffice for now.

Since the observed correlation would be expected on these favored versions of epiphenomenalism, the general epiphenomenalist hypothesis would fare better than the evolutionary argument contends once the evidence had been taken into account. If there were similar points to be made with respect to all the various correlations, or if the metaphysically favored versions of epiphenomenalism were favored enough, then epiphenomenalism might not be disconfirmed at all (or at least very little), even granting the proponent of the evolutionary argument his basic strategy.

Although this worry is an interesting one which would cause the argument trouble if it were correct, I don't see any reason to suppose that it is correct. Although it is plausible to suppose that the intensity of neural activity would be reflected in the intensity of phenomenology (e.g., small amounts of information carried to the brain from a few slightly damaged nociceptors in a limb might give rise to modest pain, while large amounts of information carried to the brain from many severely damaged but still operative nociceptors might give rise to intense pain), I can't see any grounds for supposing that something about the neural basis would predispose the likelihood of the bridge laws endowing the phenomenology with its characteristic positivity or negativity. Ultimately, though, I suppose this is largely an empirical, neurophysiological question. Perhaps there is a deep neural connectedness between the neural bases of INPs that makes the epiphenomenalist suggestion plausible. But at the moment, I'm quite skeptical, and as far as I know no scientific evidence or introspective data to this effect has been discovered.

II. The No INP/IPP Concept Objection

Now that we have examined the first three objections to the evolutionary argument and found them all wanting, let us continue on to objection (D), by far the most complicated objection we have yet encountered. This is the objection that claims we do not in fact have the concepts INP or IPP, or at least that we have good introspective grounds for supposing that no qualitative event satisfies them even if we do have the concepts.

With this objection, it will be wise to consider a specific version—that of William Robinson, one of the most prominent defenders of epiphenomenalism against evolutionary arguments, and indeed in general. Examination of it will require some lengthy and detailed exegesis and commentary, and will involve touching on some issues that will be explored in more depth later. In the end, I will conclude that it too is flawed, though showing why will be much more involved than it has been with previous objections.

Robinson on Pain and Pleasure

I will begin by examining Robinson's account of what it is to "like" a sensation (sensations are a species of qualitative event or at least components of qualitative events)—i.e., roughly, to find it pleasant—and analogously to "dislike" it. (Though he spends the majority of his time on the positive case, and so consequently we will as well, he intends the account to apply to both positive and negative cases.)[59] Robinson doesn't employ the terminology of 'INP' or 'IPP', but I think we can adapt his remarks and arguments to our current discussion without too much trouble. His fundamental claim is that liking a sensation consists in having what he calls a "meta-sensation" directed upon

[59] This account is presented in Robinson (forthcoming a)

it, where meta-sensations are conscious occurrents (presumably conscious occurrents are the same thing as qualitative events), but which differ in a variety of ways from ordinary sensations, and so merit a different classification.

After I have explored his "liking" account, I will show how Robinson criticizes the evolutionary argument, and how the account would need to be put to use as part of the criticism. Ultimately, I hope to show that in order for the account to cohere with his criticism of the evolutionary argument, it must be understood as a denial that what we would ordinarily call 'pain' is intrinsically negative, and 'pleasure' intrinsically positive. I also hope to show that it is an implausible denial at that, and thus that the objection does not succeed. (Although I will focus on the details of Robinson's specific account of liking and also on his specific formulation of the objection to the evolutionary argument, along the way I hope to show that any criticism in the spirit of Robinson's will suffer from similar pitfalls.) As I mentioned above, I will not be able to make this case with complete thoroughness in the space allotted, but I hope to outline in considerable detail how a plausible proposal would proceed, and show that the proposal is in fact plausible.

So, let's begin with Robinson's account of liking. I won't examine every claim and argument he makes on his path to the view that liking a sensation consists of having a meta-sensation directed upon it, but I will discuss all the highlights and hopefully touch on all the potentially controversial inferences that are relevant. (Incidentally, the reason Robinson considers his analysis to be of liking rather than of pleasure, is that he thinks that in some circumstances we can like sensations that we don't find pleasant. "During bouts of depression," he says, "people may not take pleasure in much of anything, but it would still be appropriate for them to say that, e.g., they like the taste of olives. It would

not give them pleasure to eat olives now, perhaps, but in general, i.e., when they are not depressed, they would get pleasure from eating them."[60] After noting this qualification, Robinson goes on to use 'liking' and 'pleasure' interchangeably, for stylistic reasons.)

It is clear, as I will discuss momentarily, that Robinson believes whatever pleasure is, it is some sort of conscious occurrent. (It does not consist in mere dispositions to continue activities or seek things out, as Gilbert Ryle famously thought.) He rejects in turn analyses that attempt to analyze pleasure in terms of having certain thoughts, desires, or emotions.[61] He rejects an analysis of a sensation *S* being pleasant in virtue of a subject having a thought that *S* IS PLEASANT primarily because, in order for the thought to be true (on a standard correspondence theory), there would need to be an independent fact—the pleasantness of the sensation—for the thought to match. But this fact is just what we are trying to get a purchase on, and so the analysis fails.

---

[60] Robinson (forthcoming a).

[61] It is generally important that we keep the notions of EXPLANATION and ANALYSIS reasonably precise in contexts like this. 'Analysis' has an *a priori* implicature to it, but common philosophical usage (employed here by Robinson on several occasions) allows the term to be used where deep and thorough introspection is concerned, even though introspecting one's sensations is not *a priori* in the same sense as reflecting on one's concepts. (Although reflecting on one's concepts does involve introspecting some vague kind of phenomenology, hence the source of potential confusion.) Since Robinson's point here is to give an analysis of the phenomenon of pleasure in the introspective sense (at least his point insofar as it is relevant for our purposes), he must not be understood as attempting to offer a scientific explanation of pleasure in terms of entities or events that are not consciously accessible (e.g., subconscious brain states). The thoughts and desires he speaks of can only be qualitative entities or events, or at least be constituted by such events. (Here I ignore issues about externalism and content, the solutions to which could imply that some thoughts or desires are not fully constituted by anything "in the head.") Robinson contributes to the confusion somewhat by mixing in a discussion of behaviorism—a candidate for real *a priori* analysis of PLEASURE—along with the introspective theories. (Behaviorism purports to give a set of necessary and sufficient conditions for being in pleasure that can be appreciated *a priori*, just by reflecting on epistemic modal space. The analyses in terms of thought and desire aim to convince a subject that, on careful introspection—a quasi-empirical pastime—she will find these heretofore unnoticed elements in her experience, and somehow be able to discern that they are the elusive pleasantness.) Robinson also risks causing further confusion when he describes desires on one occasion (outside of the behaviorism discussion) as though they were things a subject would have to infer she had, rather than directly introspect. "… people who say they want more of a taste because they like it are not drawing a conclusion from an assumption that they will take steps to continue, intensify, or repeat having that taste."

Robinson rejects a desire analysis—that liking a sensation *S* consists in having a desire for "continuation, intensification, or repetition" of *S*—mostly because there are many possible reasons why I might desire a given sensation aside from liking it. For example, I might desire a certain taste because I like it, but I also might desire it because "… I think it's good for me," or "because my religion requires consumption of this food." (Robinson plausibly believes that we can answer in these ways when the questioner focuses on the taste sensation itself, rather than the food. Although more strained, he thinks it makes sense to claim that we desire the taste because it is itself good for us or has religious significance.) But he believes that if my liking the sensation really did consist of my desiring it, then the connection between the two would seem trivial, and other answers non-sensical.[62] So the fact that the connection is not trivial and that other answers make sense destroys the attempted analysis. (Interestingly, Robinson never considers the possibility that liking a sensation consists, not in any old desiring, but in wanting more of a sensation *on intrinsic grounds* (with *ceteris paribus* clause included), rather than desiring because of some instrumental value that it has. Certainly this does seem to be a sufficient condition for liking a sensation, and it's not obviously implausible

[62] Incidentally, here we see another threat of conflation of the *a priori* sense of 'analysis' with the introspective one. Is the issue whether or not the concept of LIKING is the same as the concept of DESIRING, or is the issue whether we can introspectively notice such a desire (perhaps unnoticed previously) whenever we have a sensation that we find pleasant? The two issues may be closely related, in that what sensations satisfy the concept PLEASANT may be dependent on the presence of various desirelike phenomenological properties, but they don't clearly seem to be exactly the same thing. We must not be misled by the fact that there is some sort of vague introspection of phenomenology going on whenever we consider whether some purported set of necessary and sufficient conditions in fact has the same extension (in all epistemically possible scenarios) as the concept we are analyzing. The process of examining epistemic possibility space is a very different one from the process of carefully introspecting our phenomenology to see whether it includes certain elements. We are especially apt to be knocked off-track in cases like this because the vague phenomenology associated with our surveying of modal space bears marked similarity to remembering or imagining ourselves in a state where we are introspecting our phenomenology, since the concepts in question are phenomenological ones. Ironically, though, it is precisely because of the close entanglement between the two kinds of analysis in this case that the discussion can get by without carefully distinguishing them.

For a similar point about the relationship between introspection and the *a priori*, see Bealer (2002), p. 74.

on the face of it to suggest that it's a necessary condition as well. Maybe the problem is that the account is thought to be uninformative—to desire something on intrinsic grounds is just to desire the thing because one likes it.)

Importantly (as we will see when we examine his objection to the evolutionary argument), Robinson plausibly rejects a behavioristic account of liking as well. He does so on the grounds that a subject's knowledge of whether she likes a given sensation is not based on any inferences from behavior (whether she tends to take steps to prolong or intensify), but is rather known directly. He states:

> This view is implausible for familiar reasons. To wit, when I like a taste, I know I like it in an apparently direct way. Others may have to look to see whether, e.g., I take another bite without grimacing, or seek evidence as to whether my continuing to eat a particular food might not be merely an effort to be polite. But my own knowledge of whether I like this taste (for example) is not based on any inference from what I take my dispositions to be. On the contrary, if I think I am disposed to eat this food again, that is because I know that I like it.[63]

So, now we have seen that Robinson clearly rejects all these varied attempted analyses of pleasure. Given that Robinson ultimately settles on an account where pleasure is a kind of conscious occurrent (which is definitely not a conscious thought or desire), the first natural question to wonder about it is: why does Robinson not believe simply that likings are just sensations themselves; why does he posit this strange kind of qualitative entity, the "meta-sensation"? For our general purposes, Robinson's unique method of ontological categorizing doesn't much matter (since we are only looking at Robinson's view for its paradigmatic elements, of which this is not one), but understanding it will be helpful in maintaining clarity and appreciating with some systematicity how he is approaching the issues.

---

[63] Robinson (forthcoming a)

His primary reason for believing that a subject's liking a sensation *S* does not consist in that subject having sensation *S* and having another "liking" sensation *P* appropriately related to *S* is phenomenological. He asks us to imagine the following:

> Suppose s1 and s2 are pleasant sensations for X, which may or may not be occurring at the same time. Then X experiences s1 + /p/, and X experiences s2 + /p/. But, I claim, there is no sensation that is present on all, or most, or even several occasions on which we have a pleasant sensation. That is, there is no sensation that can play the role required of /p/. There is no sensation such that it is always, or generally, or even often present when I am having a pleasant sensation. When I taste an olive, there is the taste and it is pleasant, and when I see an expanse of International Klein Blue there is a color experience that is pleasant, but there is no sensation that occurs when I taste an olive and when I see an expanse of International Klein Blue. There is the common word, "pleasant" (or, "liked") that applies to these (and many other) cases, but there is no sensation to which this word corresponds.[64]

On the face of it, Robinson's phenomenological claim is ambiguous between "there is no sensation *simpliciter* that accompanies each pleasurable experience" and "there is no (joint-carving) sensation (type) these pleasurable experiences have in common, though there is a certain accompanying pleasure sensation (directed on the first-order sensation) for each pleasurable experience. (I.e., there is no joint-carving sensation type membership that the higher-level pleasure sensations have in common.)" It would be reasonable to suggest that a more natural reading of the passage is the latter, weaker interpretation, since he emphasizes that there is no sensation that is "always, or generally, or even often present when I am having a pleasant sensation," and because in his olive/International Klein Blue example he says that "there is no sensation that occurs when I taste an olive *and* when I see an expanse… (emphasis added)" Moreover, this kind of claim seems independently plausible. It would certainly be hard to believe that the pleasantness in the one case really is the very same kind of sensation as the

[64] Robinson (forthcoming a)

pleasantness in the other. After all, the two experiences occur in completely different sensory modalities—why suppose the pleasure-making sensation that is part of each is of the same type?

But later Robinson goes on to introduce a phenomenon he calls the "appearance of necessity"—that it is normally impossible for us to conceive of a sensation that is actually pleasant being the very same sensation, but unpleasant. He uses the appearance of necessity as the centerpiece of an argument to show that the former reading is actually the case (which seems to imply also that it is the correct interpretation of the earlier passage). The idea is that if there really is a distinct pleasure sensation directed on any given sensation that makes that sensation pleasant (regardless of whether there are commonalities between that pleasure sensation and others), we should be able to conceive of the two separately, since they are "distinct existences," to use the Humean terminology. But we cannnot conceive of them separately, so that is a good indication that there is no distinct pleasure sensation.[65]

Oddly, Robinson never explains why his preferred explanation for the phenomenon of pleasure, the meta-sensation (which is itself a kind of conscious occurrent/qualitative event which is distinct from the sensation it is directed upon) doesn't fall prey to this same sort of objection. In any case, the objection might naturally lead one to suppose that pleasure is an aspect or component of the sensation proper, rather than something outside it.

---

[65] Interestingly, Robinson claims that the appearance of necessity is mere appearance because he doesn't think "we have a good reason to think that a taste, color, etc. that a person finds pleasant must always remain pleasant for that person." Nevertheless, he believes that the appearance of necessity can do work as part of his argument against the claim that pleasure is a sensation. I am not completely clear on why this is. In any case, the important point is that he doesn't seem to believe that any pleasures are sensations, not just that there aren't important common type memberships between various pleasure sensations.

Again curiously, Robinson never gives an argument against the view that the liking is not an aspect or component of the sensation proper, rather than some sort of "add on." For example, he never gives any sort of reason for supposing that the pleasantness of the olive taste he seems to enjoy so much is not a component of the taste sensation itself, rather than a separate event or entity.

Perhaps the reason is that components of sensations are just as independent and changeable as outside conscious occurrents that are directed on sensations, and so locating the pleasantness within the sensation itself is no help. Robinson does say that "[a]lthough the components of tastes are in some sense intermingled, we can imagine less salty soy sauce, i.e., we can imagine the other taste components with less saltiness. We can often imagine what a dish would taste like if it had more rosemary in it, or if it were sweeter, and so on." He then goes on to point out that the pleasantness cannot be independently changed in these same ways.

Another possibility (and one that will become more relevant when dealing with the next objection to the evolutionary argument) is that Robinson is taking account of neuroscientific findings to the effect that pleasure phenomenology is a result of processes in a different part of the brain than saltiness phenomenology, sweetness phenomenology, etc. This may motivate his desire to treat the phenomenological contributions as parts of separate conscious occurrents.

In any event, Robinson's favored conscious entity for analyzing the phenomenon of pleasure (the meta-sensation) is never described directly owing to its fairly ineffable nature, but rather explained and illuminated by a series of metaphors and analogues—what he calls "comparison cases." The meta-sensations are "conscious occurrents" in his

terminology ("qualitative events" in ours) that in some sense constitute evaluations of the sensations they are about. And they are indeed *about* their associated sensations, in the way that feelings of finding a face familiar are about a particular face, though in a sensational and aspectual kind of way. (My pleasurable meta-sensation that is directed upon a 1986 Sangre de Toro wine taste is not directed upon it in virtue of the wineishness of the taste, but rather something more specific about it, though the taste is wineish.)

The meta-sensations are allegedly not evaluations in the sense of representations that can be true or false, though, just as the seeming familiarity of a face cannot be true or false. "What *would* be a mistake," he says, "would only be a judgment that I might go on to make on the basis of the seeming familiarity, to the effect that I had seen the face before."[66]

In any case, there is no need to dwell on these points for too long. Again, although the details are interesting in their own right, for our purposes the important part of Robinson's account, likely to be shared by most reasonable proponents and opponents of epiphenomenalism alike, are that pleasures (and pains) are qualitative events of one sort or another (whether parts of the sensations they are in some sense about, or directed on them from outside). In addition, they are somewhat more akin to sensations than to phenomenal judgments (occurrent beliefs, thoughts, etc.) or desires. (Though we will see that the similarities may not be complete.)

Although we have now obtained a fairly detailed picture of Robinson's views about concepts like PLEASURE and PAIN (the concepts that would be candidates for species membership under the broader IPP and INP concepts) and the introspectively accessible phenomena of pleasures and pains, we have not explored in any detail the

[66] Robinson (forthcoming a). Emphasis in original.

issue of whether in fact Robinson's account does have a place for things like intrinsic positivity and negativity.[67] In order to get a better handle on that question, we will need to proceed on to examine his objection to the evolutionary argument, contained in a separate (but also very recent) paper.[68]

Robinson's Objection to the Evolutionary Argument

Recall that Robinson is one of the most prominent defenders of epiphenomenalism today, and in particular, against evolutionary arguments. In his paper "Evolution and Epiphenomenalism," he formulates a version of the evolutionary argument and ultimately rejects it precisely because he thinks we do not have something akin to the notions of INP or IPP (or at least no introspective reason to suppose anything falls under those notions).

Because consideration of the evolutionary argument and the quality of phenomenology is just one part of his paper, the formulation he considers is not as detailed as the one set out earlier in this dissertation and is laid out completely informally. For that reason, there is no point in examining it in depth. The basic framework of the argument explicitly owes its inspiration to William James' version from *The Principles of Psychology*.[69] It asserts that there is a "hedonic/utility match" that stands in need of explanation, and that the only plausible way to explain that match is scientifically. Moreover, the only plausible scientific explanation is evolution, and the only way that

---

[67] In the discussion coming up, I do not attempt to pin down in detail what the phenomena of introspection and introspective judgment amount to. A high degree of precision is not required, and might actually detract from the discussion by introducing issues irrelevant to the central question at hand. Later, in connection with issues about introspection and epiphenomenalism, I will consider substantive issues surrounding introspection in a bit more detail, though I will still have to set many substantive issues aside for further research.

[68] See Robinson (forthcoming b)

[69] James (1890).

evolution would select for the match would be if pains and pleasures were efficacious.[70] (It is interesting to note that this argument focuses on a match between phenomenology and behavior rather than phenomenology and distal stimulus. But I don't think there will be any major differences in outcome as a result of this shift in emphasis, as will be discussed more fully later on, in the next chapter.)

The huge issue for Robinson (the same issue that we are exploring in this chapter) is what (if anything) is wrong with this argument that, if correct, threatens to bring the epiphenomenalist view tumbling to the ground. Robinson is crystal clear (as most sensible individuals, and all epiphenomenalists would be) that the hedonic/utility match cannot be analyzed away behavioristically, by claiming that "'pleasant' just means "what we generally prefer without coercion", and 'painful' just means "what we generally avoid without coercion."[71]

With this easy way out denied, the goal for the epiphenomenalist is two-pronged—first, explain how the 'because' in claims like "I ate the olives *because* I like the taste of them"[72] is not a causal "because," and second, avoid claiming that pleasantness or painfulness consists "simply in the fact of being pursued (avoided) without coercion."

Meeting the first challenge is easy, since the epiphenomenalist can simply claim that the relevant notion of BECAUSE is one of counterfactual dependence, not causation.

---

[70] It is worth noting that the version of the argument Robinson considers, like most traditional formulations, pits epiphenomenalism against both interactionist dualism and physicalism. Because I will be considering the question of where interactionist dualism fits into the picture at length in a later chapter, I have deliberately restricted the discussion here to an argument that the evidence confirms physicalism over against epiphenomenalism. Fortunately, Robinson's treatment focuses primarily on issues that have little or nothing to do with interactionism's place, and so his inclusion of interactionism on the physicalist side can be ignored.

[71] Robinson (forthcoming b)

[72] This is my example, not Robinson's.

And it is clear that, according to epiphenomenalism, my eating the olives will be counterfactually dependent on the liking. The reason is that, if the liking were not present, neither would the eating of the olives, since the closest worlds where I did not like the taste would be worlds where the neural base of the liking was removed, and it is very plausible to suppose that this neural base is causally efficacious in my eating the olives and would not be causally replaced by something else.

Meeting the second challenge is much more difficult. Robinson sketches out the beginnings of a proposal of how this is to be done:

> Let us suppose that an organism, O, does some action, A, in circumstances, C, and that the world produces a result, R upon O's body. Let us suppose further that R causes an experience of a certain kind, E, and an instance of N(P)… N(P) will have two kinds of effect. (a) It will cause P. It is a further, nontrivial fact that this P will be directed upon E. Here, I will not go into the explanation of this fact; I will just assume that the neural causes of P and of E are related in some special way that regularly underlies the directedness of pleasure onto the experiential quality. (b) N(P) will cause the motivational system to raise the probability of doing A again when circumstances similar to C are encountered…. Now, let us suppose that doing A in C is beneficial to O. Then a consequence of the scenario just outlined will be that O will tend to repeat the beneficial action, and will find the results pleasant. Suppose further (1) that O's brain is organized so as to increase the robustness of the connection between the beneficial action and the occurring of N(P). Then O will become more likely to repeat the beneficial action. Suppose, however, (2) that O's brain is organized so as to reduce the robustness of the connection between the beneficial action and the occurring of N(P). Then O will become less likely to repeat the beneficial action. The combination of these two points licenses the conclusion that selectional pressures will favor organisms like those in (1) preferentially to organisms like those in (2). That is, the connection between beneficial action and occurrence of N(P) can be favored by natural selection.

Although the above passage does make clear how natural selection *could* select for a given pleasurable experience (composed of sensation E and pleasure meta-sensation P) even if that pleasurable experience were not causally efficacious in any behavior, it doesn't make clear why such a selection is especially likely (and *mutatis mutandis* for

pain).[73] The account here does explain how, if the experience composed of phenomenal elements E + P is caused by or emerges from a certain neural basis via some law of nature, that experience could be selected for—namely, by being a mutual effect, along with advantageous behavior, of the neural basis. But, as our original formulation of the evolutionary argument above made clear, this approach leaves unexplained the *prima facie* amazing coincidence that negative phenomenology is caused by neural events that produce aversive behavior, and positive phenomenology by neural events that produce seeking behavior.[74]

Robinson is well aware of this shortcoming in his account, and aims to rectify it. He speaks of a "disturbing thought" that might occur to us at this point: "What if the neural event that has been labeled 'N(P)' were to have caused displeasure instead of pleasure? It doesn't seem that natural selection could get a handle on *that* mismatch. But, if that is right, then aren't we faced with the Jamesian spectre after all? Must we not say that, if epiphenomenalism is assumed, natural selection cannot get a grip on maintaining hedonic/utility match in preference to mismatch?"[75]

He responds by contending that this worry is ultimately incoherent, but by sliding between competing incoherent possibilities, we can convince ourselves that it is coherent. He gives the following two incoherent cases, his presentation of which should be given in its entirety:

> Case A. One imagines an internal suffering that cannot be given overt expression. Poor Jones is inwardly railing at the disgusting taste he's suffering, even as he stuffs more and more olives into his mouth.

---

[73] In this paper, Robinson never explicitly endorses the meta-sensation account, preferring instead to remain officially agnostic about the nature of this P, aside from its being conscious. I bring in the meta-sensation account because it is motivated by his criticisms of some alternatives in "What is It Like to Like?," and shares what I believe to be the relevant features with any rival account for present purposes.

[74] This is a coincidence that will receive more discussion in later sections.

[75] Robinson (forthcoming b)

> One can indeed imagine a bizarre case of this kind by imagining simultaneous occurrence of locked in syndrome (in which one can subvocally say things to oneself but not initiate any motor effects) and a seizure that repeatedly sequences the motion of picking up an olive and transporting it into one's mouth. But it is incoherent to suppose that tendencies toward occurrences of this sort could not be selected against. Such a supposition is incoherent because the occurrence is a breakdown in behavioral organization, with the products of cognition unable to affect action. The supposition amounts to saying that tendency toward a disease cannot be accessed by selectional pressure. That is incoherent, and that is why it is incoherent to suggest Case A as a way of cashing out the Jamesian spectre.
>
> Case B. One imagines perfect equanimity, with actions and subvocal speech just as they normally are. The only difference from our case is that instead of N(*P*) causing *P*, it causes something else – let us say X. Maybe X is the same as what we would call displeasure, or maybe it is something else entirely.
>
> This case is also incoherent, but for a quite different reason. Namely, the only way we have of identifying which conscious occurrence is *pleasure* is that it is the one that goes with activities we tend to repeat without coercion. Whatever X this is (assuming there is one) is the conscious occurrent that is referred to by "pleasure." Consequently, it is incoherent to suppose that X is the conscious occurrent that typically goes with activities we pursue without coercion, but is not pleasure.

So the difference between the two cases is that in Case A, there are lots of internal (i.e, phenomenological) processes like subvocal sayings going on, but that no signs of these processes can manifest themselves in overt behavior, while in Case B, all the internal processes are just as they normally are, but P has changed to some other conscious occurrent, which we would naively describe as negative.

Before continuing on to consider an objection Robinson raises to his claim that Case B is incoherent, let me pause for a moment to ensure, for the record, that I deal clearly with what Robinson calls "subvocal sayings." It certainly sounds like subvocal sayings are phenomenal verbal imagery, and henceforth I am going to assume that they are. If they are something else (like tiny inaudible movements of the vocal chords or the antecedent firings of various extra-cortical neurons), then it clearly will be incoherent to

claim that they cannot initiate motor effects or other overt consequences. But it will be implausible to suppose they necessarily accompany the internal suffering, because there is no reason to believe that the neural correlates of that suffering would have to be even causally connected to the motor effects. (*Ex hypothesi* the subvocal sayings in Case A genuinely have no way of making their presence felt in behavior, either as a result of causing the behavior or the behavior being counterfactually dependent on them. Consequently, whatever neural event causes them must not be causally active in producing any of these vocal chord movements.)

The objection Robinson considers to his incoherency claim is that he is, after all, "reneging on [his] commitment not to reduce pleasure to uncoerced repetition." His reply is that he is not asserting that PLEASURE is equivalent to SEEKING WITHOUT COERCION or the like, but rather that he is fixing the reference of PLEASURE by "invoking the role of being caused by activities we repeat without coercion."[76] (I.e., PLEASURE functions as a *de jure* rigid concept.) He explains what he believes is an analogous case, that of WATER, and how it illuminates the phenomenon:

> …the reference of "water" is fixed by invoking the role of being the compound that composes the contents of lakes. But in a nearby possible world in which lakes are filled with XYZ, they are not filled with water. Just so, I suppose that there is a nearby possible world in which the psycho-physical laws are different, and the same neural event, N(*P*), causes X instead of *P*. Residents of that world will call X "pleasure", but the right thing for *us* to say, on the view I am proposing, is that pleasure is *P* and that what they call "pleasure" is not pleasure, but X.
>
> One may at this juncture imagine that X is what we would call "displeasure", and proceed to imagine the residents of the imagined possible world as suffering as they lustily gobble their food. But that would be to return to the incoherent Case A. It is the ease of making this transition that, I believe, ultimately accounts for the plausibility of James's argument.

[76] It is not totally clear from Robinson's notation whether he intends the discussion to be about terms (which are linguistic entities, obviously) or concepts. Since the original argument was about phenomenal concepts rather than terms, and because the discussion up until now has focused on concepts, I will interpret Robinson to be discussing concepts here.

Now that we have set out in detail Robinson's objection to the evolutionary argument, it is time to show what is wrong with it—namely, that it requires a view of pleasure and pain whereby the corresponding meta-sensations are not intrinsically positive and negative respectively, and that this denial is implausible.

Some readers may find Robinson's suggestion that concepts like PLEASURE and PAIN refer rigidly only *de jure* too counterintuitive to take seriously, but to this response, I have two comments. First, even if the suggestion is extremely counterintuitive, just as with epiphenomenalism itself, it would be nice to say something more substantive than simply that it is counterintuitive. And second, there is more at stake here than simply whether these concepts refer in the way Robinson thinks—there is also the issue of intrinsic positivity and negativity. Although I think it is evident that defeating Robinson's claim is an important step in clearing the way for the notions of INP and IPP, it is not equivalent to showing that INP and IPP are coherent after all. More must be said to combat the overall objection being discussed in this section of the paper—i.e., that we do not have the concepts INP or IPP, or at least do not have introspective grounds for supposing anything actually falls under them.

I would first like to focus directly on Robinson's account of meta-sensations and pleasure/pain, and how this account relates to the *de jure* rigid conception of PLEASURE he introduces in "Evolution and Epiphenomenalism." In the process, I hope to show that our concept PLEASURE (and PAIN *mutatis mutandis*) doesn't function in the way Robinson supposes in considering these cases, and moreover that supposing it does undermines the primary motivations for the account he develops in "What is it Like to

Like?," and generally for any similar account that gets its force from the same reasonable considerations. Subsequently, I will apply these results to a discussion of Case A and Case B in turn, in an attempt to demonstrate that there are coherent possibilities in the vicinity after all, especially when we conjoin the cases with a plausible view of the phenomena of pleasure and pain (and in fact one in the basic spirit of Robinson's own in "What is it Like to Like?").

The Problems with Robinson's Conception of PLEASURE and PAIN in His Response to the Evolutionary Argument

The key claims from above are: (1) what it is for an experience to be pleasant is for it to have P (as we know from "What is it Like to Like?", P is a meta-sensation) directed on the sensation(s) that are part of the experience. And (2) P cannot be entailed by any behavioral facts (or presumably facts about dispositions to behave). The reason is that if pleasantness (of the sensation) just consisted of having a disposition to seek out it or its causes uncoerced (or consisted of the actual uncoerced seeking out of such things), then there would be a true conditional claim of the form "if [insert behavioral or dispositional fact here involving the seeking out of the cause of the sensation], then the sensation is pleasant," holding with the strongest modality. But, according to Robinson there are no such conditionals holding with the strongest modality, otherwise behaviorism would be true—it would be a necessary truth (and presumably an *a priori* accessible one) that in every possible world, the disposition to seek out would be accompanied by pleasure. If behaviorism were true, though, the second of his two requirements for defeating the evolutionary argument would be violated. (Recall that the two conditions

were first, that epiphenomenalists be able to make sense of the claim that we seek things out *because* they give us pleasure, and second, that PLEASURE not be analyzable as SENSATION WE ARE DISPOSED TO SEEK OUT THE CAUSE OF.)

Later, of course, as we have seen, Robinson will develop his account further and claim that although there are no true conditionals of the above form that hold with the strongest modality, there are true material conditionals of the above form. In other words, it is true that in the actual world, if certain behavioral facts hold, there always is an accompanying pleasant sensation. Moreover, he claims that it is by learning words that have their reference rigidly fixed by appeal to these behavioral facts that we form our concepts for classifying sensations as pleasurable and unpleasurable. According to Robinson's theory, it looks like our only methods for classifying sensations as being pleasurable or unpleasurable (i.e., of having the appropriate meta-sensations directed on them) are via their connections to certain kinds of behavior and behavioral dispositions. Even if there is a special kind of common intrinsic character associated with the various meta-sensation types, or the complexes of regular sensations and meta-sensations (e.g., burning sensation with appropriate displeasure meta-sensation), our only way of tracking the commonalities in them is via these connections with behavior and behavioral dispositions.

The issue I want to explore now in more detail is whether or not pleasures in fact do all have anything interesting intrinsic in common.[77] (Two entities have something interesting intrinsic in common iff they share some relevant intrinsic property in common that "carves nature at its joints" (or close to its joints), to use the common Platonic

[77] Since Robinson focuses exclusively on pleasure in this part of his paper, I will do the same in my discussion. It should not be difficult to apply the results to pain.

metaphor. In this case, that something interesting would have to be more substantial than just that they are both qualitative mental properties, for instance.)

At this point, I think Robinson's account faces a dilemma. Either there is some deep, intrinsic similarity between pleasures (or at least some large subset of the pleasures, such as all the bodily pleasures), or there is not. I will argue in this subsection that if there is not, then Robinson must endorse a number of claims that are both implausible on their own merits and also in conflict with the spirit of his "meta-sensation" account (in conflict with the spirit of any account that aims to plausibly explain the same kinds of general considerations, in fact). If, on the other hand, there is some deep intrinsic similarity, then that will set the stage for an argument in favor of the coherence of at least one of the two cases A and B that he claims are incoherent, and a vindication of the notions of an IPP and an INP (since intrinsic positivity and negativity are plausible candidates for being one of the interesting common things in the respective classifications, or the single interesting thing). Momentarily, in connection with the dilemma, we will see how the view of PLEASURE as *de jure* rigid is problematic.

I'll begin by considering the second horn of the dilemma, since some textual evidence suggests that it is the one Robinson actually endorses. (Hereafter I will refer to this horn of the dilemma as the NIIC horn—short for "nothing interesting intrinsic in common.") For instance, take his paradigmatic claim that "the only way we have of identifying which conscious occurrent is *pleasure* is that it is the one that goes with activities we tend to repeat without coercion. Whatever X this is (assuming there is one) is the conscious occurrent that is referred to by 'pleasure'." If pleasures really did have something interesting intrinsic in common, then it seems plausible to suppose that we

could identify pleasures that way, rather than via their connection with behavior or behavioral dispositions.

As promised a moment ago, I'll explore the problems with the *de jure* rigid conception of PLEASURE—the view that claims that the concept picks out qualitative events via their accidental association with the right sorts of dispositions—since the lessons we learn are directly applicable to showing why it so implausible to suppose that pleasures do not have anything interesting intrinsic in common. To appreciate the arguments, however, *it is critical to think of this PLEASURE concept as the one we actually employ in introspective judgment (and perhaps thought generally), not as the one corresponding to the word of public language and providing its meaning.* I have no qualms with the suggestion that we communicate with one another using words that specify mental events via their connections with behavior and behavioral dispositions. (How else would we efficiently communicate except by making use of the realm of things that are readily intersubjectively accessible?) But I do have qualms with the suggestion that in this case we introspectively think and classify using concepts similar to the ones we use in public communication.[78]

I have three fundamental arguments against the *de jure* view, separate but related in a number of ways (especially the first and third). The first is based on the very simple insight that we do classify qualitative events as pleasures introspectively. (At no point in considering the present issue with the evolutionary argument will I call into question epiphenomenalism's ability to account for introspection and introspective judgment, or

[78] For a similar view, see Chalmers (2003). There, he claims that "we do not have public language expressions that distinctively express the content of [pure] phenomenal concepts."

the coherence of introspection on epiphenomenalism. Those are difficult issues, and will be dealt with much later.) The argument goes as follows:

(A) A subject is able to introspectively classify a qualitative event iff he is able to classify it without taking into account anything that presents itself to him as being a fact about the external world or his behavioral dispositions.

A word about premise (A). The "presents itself" qualification is meant to protect against the objection that qualitative events may count as facts about behavioral dispositions, because the objector believes that they ground those dispositions or somehow implicitly encode information about them at the subconscious level. If (A) simply read "… classify it without taking into account any facts about the external world or his behavioral dispositions," the objector would claim that the facts accessed in introspection were facts about behavioral dispositions, even though they did not seem like such to the subject. For an agent to "take into account" something is for that agent to use the thing for evidence, or for the thing to cause the classificatory judgment in the right sort of immediate way.)

Premise (A) is intended to be a direct consequence of an analysis of INTROSPECTION.

(B) If an agent is able to classify a qualitative event without taking into account anything that presents itself to him as being a fact about the external world or his behavioral dispositions, then he is able to classify it based on its intrinsic

features.

So, from (A) and (B):

(C) If an agent is able to introspectively classify a qualitative event, then he is able to classify it based on its intrinsic features.

(D) Agents are able to introspectively classify pleasures (as pleasures).

(D) is a fairly obvious assumption, and Robinson readily makes it. That is apparent in "What is it Like to Like?," when he says—"To wit, when I like a taste, I know I like it in an apparently direct way. Others may have to look to see whether… I take another bite without grimacing, or seek evidence as to whether my continuing to eat a particular food might not be merely an effort to be polite. But my own knowledge of whether I like this taste… is not based on any inference from what I take my dispositions to be. On the contrary, if I think I am disposed to eat this food again, that is because I know that I like it." This "apparently direct way" of knowing one likes a taste must be introspection.

From (C) and (D):

(E) Agents are able to classify pleasures based on their intrinsic features. (Note that it could be argued that the passage cited in connection with (D) could

potentially get us to this premise directly, thus bypassing (C) and (D). The motivation for the claim is clearer and less controversial with the supporting argument, though.)

It is difficult to dispute the soundness of this argument. (A) is a fairly uncontroversial analysis of INTROSPECTION (at least in its relevant details), and (D) is very attractive and also clearly endorsed in "What is It Like to Like?" The only other assumption, (B), is a highly plausible general claim. It can be challenged (as I will explore below), but challenging it involves committing oneself to implausible views of how introspection and introspective judgment work that anyone defending the NIIC claim will also be driven toward. In any case, even if the *de jure* rigid view of PLEASURE can escape defeat at the hands of this argument by clinging to those implausible views, it will be discredited by one of the later arguments.

The conclusion of this argument is manifestly inconsistent with the *de jure* view, since that view claims that we classify and designate qualitative events based on the accidental extrinsic feature they have of being associated with certain behavioral dispositions (in fact, not just that we *do* classify them this way, but that that is all we are able to do). Consequently, that conception of PLEASURE must be rejected if the argument is correct. Just in case there are readers who do dispute the soundness of it (in particular, by disputing (B)), as I said above I have other arguments against the view that PLEASURE functions in the way Robinson supposes. Before considering them, though, I'd like to consider the relevance of the conclusion of the present argument for the slightly bigger picture issue of whether the NIIC view is correct, and in the process

discuss the prospects of rejecting (B).

In and of itself, the conclusion is consistent with all the pleasures failing to share something interesting intrinsic in common. The problem arises for the NIIC view when we consider what introspective judgment would have to be like if it classified pleasures based on their intrinsic features, but where those intrinsic features were not unified in any interesting way.

A big problem for reconciling the conclusion of the argument with the claim that pleasures have nothing interesting intrinsic in common is that, if the NIIC view were in fact correct, it seems introspection about pleasure would be much more highly labored than it is. The problem is bad for pleasurable experiences exactly resembling old ones, but it is even worse for pleasurable experiences that are brand new and unprecedented.[79] Let's consider the case of new tokens of familiar types first. In order to identify such a qualitative event as a pleasure, people would have to (by some unconsciously directed process)[80] recall all the various kinds of phenomenology associated in the past with dispositions to seek out (hereafter, SODs). Since *ex hypothesi* these qualitative events would have no interesting intrinsic features in common, there would be no organizational principle or readily available procedure that would allow for speedy and efficient classification of a currently experienced qualitative event as a pleasure. But plainly we

[79] There may be difficulties for epiphenomenalism in making sense of how someone could episodically remember a qualitative event, since such memory would seem to inextricably involve causation on the part of the qualitative event. It may be possible for the epiphenomenalist to understand the role involved as a specific kind of counterfactual dependence rather than outright causation, but the issues such an account raises will get tricky. Although interesting, the worry is too far off topic for me to pursue further here, though it should be examined by anyone ultimately interested in making a case for epiphenomenalism. I am grateful to Brian McLaughlin for pointing it out.

[80] Consciously recalling information about dispositions as part of the introspective classification would violate the analysis of INTROSPECTION in (A), since it would involve using information in the classification that presents itself to the subject as information about dispositions.

are able to speedily and efficiently classify qualitative events as pleasures, which suggests that introspection and introspective judgment work in some other way.

In fact, if the brain did a poor job of linking memories of qualitative events with the dispositions associated with them, introspective attempts to classify new qualitative events (but qualitative events exactly similar to ones had previously) as pleasures might not just be inefficient, but an utter failure. The best an agent could hope for is that her psychological architecture reliably organized stored qualitative memories according to the dispositions associated with them (SODs in this case). (A toy picture of how this might work is to imagine the brain as containing compartments for qualitative memories. One compartment would be devoted to qualitative memories that were accompanied by SODs—with subcompartments perhaps for determinate kinds of SODs. When a new qualitative event occurred, the brain would compare its overall qualitative character—i.e., combined sensory and metasensory character—with the stored memories. If it found a perfect match with a qualitative memory in the SOD compartment, it would automatically classify the event as pleasurable. This is the main way (B) could be challenged in fact—by claiming that this subconscious way of sorting qualitative events, although ultimately working by finding qualitative matches in the "memory banks" for new qualitative events, still made the classifications by noting what dispositional category the matching memory was located in. It could then be contended that any procedure that takes into account this kind of information is not classifying based on intrinsic features, though it is classifying introspectively.)[81] At least this setup would

[81] Note that this method doesn't violate the above analysis of INTROSPECTION precisely because, although the comparison procedure would employ information about behavioral dispositions in the classification, this information would not present itself as such to the subject. (The subject would not have conscious access to the real nature of this comparison procedure—it would just seem like an automatic

allow the search to focus confidently on one subset of qualitative memories, but it would still have to involve a one-by-one comparison between the new qualitative event and the various stored ones, which would inevitably make introspection far slower and clumsier than it actually is. (The agent could really luck out and be able to get away with some kind of family resemblance comparison procedure. This would work if some pleasures had common features with some other pleasures and only other pleasures, but not with all. Even so, though, the number of different "families" would likely be large, and so the advance in efficiency only minimal.)

As I mentioned above, the problem is even worse for pleasurable experiences unlike any we have experienced previously. How would we be able to introspectively recognize, for example, that the experience when tasting a never before sampled entrée was pleasurable? (And plainly we are able to introspectively recognize such things. When someone asks us if we like a food we have just tasted for the first time, for instance, we don't feel the need to say something like "I'm not sure, bring the food around again and I'll see if I'm disposed to take some more.") There would be no paradigmatic episode to compare the experience to, nor any simple algorithm for the mind to perform in deciding whether the properties of the experience are of the right sort to qualify it as pleasurable. To return to our metaphor, there would be no relevant qualitative memory in the "SOD compartment" to serve as the matching entry.

A suggestion designed to alleviate these kinds of problems would be that currently experienced pleasant qualitative events are "attached" to a record of the

---

"hunch" at the conscious level.) Also, see the note just above on qualitative causation. The way I have described the process here seems to violate epiphenomenalism, since (e.g.) "comparing" qualitative characters seems to involve qualitative causation unacceptable on epiphenomenalism. In order for epiphenomenalists to make this talk coherent, they will have to paraphrase the objectionable descriptions of the processes involved.

dispositions associated with them, and so introspection can access information about dispositions, and the agent can efficiently classify qualitative states as pleasures purely introspectively, purely by appealing to behavioral/dispositional facts, and without supposing those pleasures had anything interesting intrinsic in common. (This might also be a way of denying (B), when combined with the claim that this sort of introspective judgment is really not classifying based on intrinsic features.) The attachment would have to be carefully explicated to ensure that it did not involve a violation of the above analysis of INTROSPECTION (and included, say, appealing to observations that presented themselves as being about behavior or behavioral dispositions). The most plausible way to make the suggestion work would be to claim that the dispositional record is somehow embedded in the phenomenology of the qualitative event itself and a suitable analogue in the underlying neural basis of the qualitative event. (In fact, as far as I can see, this is the only minimally plausible way to make the suggestion work.) This theory would differ from earlier suggestions in that the dispositional information would not be accessed based on the location where memories qualitatively identical to the current experience were stored. Rather, the information would somehow be directly a part of the phenomenology.

My response is simply to deny that there is any such encoding of dispositional records in phenomenology or in its underlying neural correlates. (This is especially true for epiphenomenalism, since the feature of the view that got it into trouble to begin with was the lack of need for any fit between phenomenology and behavior. But even on other mind-body theories, just because phenomenology or its neural basis is the ground of a disposition wouldn't imply that it contained the right kind of record of dispositional

information to be employed in this kind of introspection.) I'm not sure what else can be said—it just seems undeniable that introspective classification works in some other way. Certainly, there is no explicit awareness in introspection or introspective judgment of classifying events by the dispositions associated with them. And it would be truly shocking if what is really going on in introspective judgment is a subconscious classification process performed based on dispositions associated with and informationally embedded within the qualitative events, in spite of all conscious appearances that that is not what is going on. How would we fail to notice the phenomenal components in question?

To sum up this discussion, obviously introspective classifications of qualitative events as pleasures works far more quickly, efficiently, and accurately than the picture being examined would suggest, so there must be something wrong with the picture. (In fact there are further problems with introspection on this picture that will come up once we have examined some of the scientific findings regarding the relationship between different parts of the brain in generating pain and pleasure phenomenology. I will postpone presentation of these until that information has been discussed.)

Now that we have seen the first argument against the *de jure* rigid view of PLEASURE, let's consider the second of the three. It is also quite simple, though lacking in the same kinds of systematic consequences for the NIIC view. The basic problem is that SODs actually occur without pleasure in some cases, and perhaps even pleasure without SODs—but this fact is incompatible with the analysis of PLEASURE that claims what are picked out are the experiences caused by the neural grounding in the actual world of all SODs. (According to epiphenomenalism, this must be how PLEASURE

refers on this basic view of the concept; it could not rigidly designate the grounding of the disposition itself, as physicalism might contend, since intuitively this would pick out the wrong kind of thing—a non-qualitative brain state rather than an experience.) Robinson, in fact, all but acknowledges these cases in "What is it Like to Like?" Recall his intuitively plausible rejection of analyses of pleasure in terms of desire, on the grounds that we can desire things for reasons other than that they are pleasurable. We can think they have religious significance, for example, or because they are good for our health. Although we were understanding these claims to apply to "desire" in the occurrent, phenomenological sense, analogous points apply to dispositions. Plainly, we are sometimes disposed to seek out activities and stimuli for reasons other than that they cause neural events that in turn cause pleasure—because we believe they have religious significance or are good for our health, for instance. The neural grounding of these dispositions do not in turn cause pleasure phenomenology (they may cause painful phenomenology, in fact), so the purported theory—which claims that PLEASURE does refer by picking out the experiences caused by these neural events—fails, since the theory implies that PLEASURE refers to things it plainly does not.

Although more unusual, there also may be cases where pleasures occur without SODs. In "What is it Like to Like?," Robinson says that "… the medical literature is replete with distressing cases of cognitive breakdowns that take the form of nearly incomprehensible dissociations." (In fact, we will look at one such form of dissociation—reactive dissociation—in a later section of this chapter.) He also suggests that "there is no way to rule out the possibility of suffering a cognitive breakdown that would dissociate our sincere reporting and other behavior from our actual liking and

disliking." (I take it that this possibility is intended to be epistemic—that it could turn out that actual people suffer such breakdowns, that we may wind up discovering that they suffer such breakdowns.)

If sincere reporting and other behavior is sometimes divorced from our genuine liking in this way, then people may be disposed to avoid things *ceteris paribus* that in fact give them pleasure. Just the epistemic possibility that this is the case—that it would be coherent for things to turn out this way—is enough to discredit the *de jure* analysis further, since the *de jure* analysis claims that PLEASURE functions by picking out the experiences that happen to be associated with SODs in the actual world. In other words, PLEASURE functions by locating all the SODs, and designating all the experiences caused by the neural states that ground the dispositions.

There may be ways to amend the account in relatively minor ways to get around this argument (by building in certain qualifications to the proposed analysis), but I don't think the same goes for the next and final argument. Just like the first one, I think that this one is highly problematic for the *de jure* view, and has serious implications more generally for the NIIC claim.

The third argument builds off of the first in some ways, and aims to show directly that pleasure must be intrinsically unified. It is a *reductio*, and goes as follows:

(1) If PLEASURE really were analyzable as QUALITATIVE EVENT I AM ACTUALLY DISPOSED TO SEEK OUT THE DISTAL CAUSE OF (where ACTUALLY functions as a rigidifier), then it would not be guaranteed that all the qualitative states picked out by the concept would share any interesting

intrinsic features in common. (The reader may add *ceteris paribus* bells and whistles to this analysis if desired.)

(2) Pleasure is not a gerrymandered kind. (Assumption)

(3) If some kind is not a gerrymandered kind, either it is guaranteed that all the members of the kind have some interesting intrinsic feature in common in all worlds where they are members of the kind, or it is guaranteed that they have some interesting relational feature in common in all worlds where they are members of the kind. (Note—'interesting' here is just shorthand for something like "fairly joint-carving" or "fairly natural," in the metaphysician's sense of 'natural.')

(4) All pleasures are essentially pleasures. (I.e., they are pleasures in every possible world where they exist.)

From (2), (3), and (4):

(5) It is guaranteed that all the members of the kind pleasure have some interesting intrinsic feature in common… or it is guaranteed that they have some interesting relational feature in common…

(6) It is not the case that it is guaranteed that all the members of the kind pleasure

have some interesting relational feature in common... (Though they may have some interesting relational feature in common in the actual world—namely, their connection to certain dispositions and behaviors—the analysis of the concept in (1) provides no assurance that those features will hold across worlds.)

From (5) and (6):

(7) It is guaranteed that all the members of the kind pleasure have some interesting intrinsic feature in common…

(8) PLEASURE is analyzable as QUALITATIVE EVENT I AM ACTUALLY DISPOSED TO SEEK OUT THE DISTAL CAUSE OF.

(Assumption for *reductio*)

But, from (1) and (8):

(9) It is not guaranteed that all the qualitative states picked out by the concept PLEASURE would share any interesting intrinsic features in common.

It is a trivial exercise to supply additional premises that make the contradiction between (7) and (9) explicit. So now what assumption to reject?

(1) seems obviously true. If PLEASURE were analyzed this way, the things it picks out could have very different intrinsic constitutions. If WATER, for example, were analogously analyzed as THE SUBSTANCE OF AN ACTUAL CLEAR LIQUID SAMPLE THAT PEOPLE DRINK OR THAT FLOWS IN RIVERS AND STREAMS, then we could not ensure *a priori* that there would be any common intrinsic feature that the various clear liquid samples would share in common. (How could we know *a priori*, for instance, that only one kind of substance quenches our thirst?) On the other hand, if WATER were really analyzed as (and most people prefer something closer to this way of analyzing it) THE COMMON SUBSTANCE IN THE CLEAR LIQUID SAMPLES THAT PEOPLE DRINK OR THAT FLOWS IN RIVERS AND STREAMS, we do ensure *a priori* that all the samples of water share some intrinsic feature in common. But, we fail to ensure that the extension of WATER is non-empty. (Since, again, we have no *a priori* guarantee that there is a common substance—specified intrinsically—that all the samples share.) Similarly, if we were to adjust the analysis of PLEASURE to ensure a common intrinsic feature among all the pleasures, we would gain no assurance thereby that there really were pleasures. We would only learn that if there were pleasures, they'd have these common features. So I conclude that it is useless to reject (1).

(3) is virtually a definition of 'non-gerrymandered kind', (4) a highly plausible assumption about pleasures, and (6) a philosophical consequence of the analysis in (1). The only remaining options are (2) and (8).

What can be said in favor of (2)? Well, for one thing (as I will address in more detail below), if pleasure were a gerrymandered kind, then it would not be readily

apparent what the motivation was for positing meta-sensations in the first place.[82] For another, it would be very difficult to diagnose people's lack of comfort with the idea (e.g.) that organisms in another world could not find qualitative events of the sort we associate with being burned pleasant, just because they happened to be hard-wired to seek out the causes of such qualitative events. If pleasure were a gerrymandered kind, then it would be hard to put much intuitive emphasis on the fact that burning-feels just happen to be associated with actual dispositions to seek out an end to them.

This point probably isn't entirely clear, so let me illustrate with an analogous case—the water example again. Imagine the analysis of WATER really was THE SUBSTANCE OF AN ACTUAL CLEAR LIQUID SAMPLE THAT PEOPLE DRINK OR THAT FLOWS IN RIVERS AND STREAMS, which is analogous to the proposed analysis of PLEASURE. And suppose the actual world had clear liquid samples with tremendous diversity of dissimilar intrinsic constitution—H_20, ABC, DEF, LMN, etc., but not XYZ. Although I can give no demonstrative argument for my claim, I suspect that the inhabitants of the actual world in this case (where the actual world here is of course relative to the example) would not think that including the clear liquid on Twin Earth, which is uniformly XYZ, under the concept WATER would be much of a departure from their present usage. Of course, the philosophically savvy among them would recognize that XYZ is not technically water, but they would regard including it as a relatively minor violation of conceptual convention, probably not worth losing sleep over and certainly fixable by relatively minor stipulative changes to the conceptual scheme. Contrast this with the strongly negative intuitive reactions people in our world

[82] I don't want to rest too much of my defense of (2) on this claim, though, since part of the motivation for it is that (2) is independently plausible. Consequently, making this claim central to the defense of (2) would open up the serious possibility of vicious circularity.

would get when it was suggested to them that Martians who are disposed to seek out the causes of the kinds of qualitative states we associate with burnings are in pleasure. People would regard this as a serious, highly non-trivial violation of our concept of PLEASURE and one that could not be fixed by a minor stipulative change here and there.[83]

Another serious problem for (2) is the introspection worries we have already seen in connection with the first argument against the *de jure* view. If pleasure really were a gerrymandered kind, then introspectively identifying pleasures would be much slower and less efficient than it is. (It is important to recognize that simply challenging the first argument by denying the only semi-controversial premise—the claim that introspective classification implies classification by intrinsic features—doesn't help to get around this result, because the current argument doesn't rely on the claim that introspective judgment makes classifications by intrinsic features.)

So, I conclude that (2) is worthy of acceptance. The only alternative left is to reject (8), which is the proposed *de jure* analysis of PLEASURE.

Now that we have seen the problems with the *de jure* view, let's return to the main task, which is examining the NIIC horn of the dilemma. A potentially serious problem (alluded to above) is that if there is not some deep intrinsic similarity between pleasures, then we must justify the positing of the meta-sensations in the first place. Presumably, the meta-sensations were posited to begin with because there seemed to be a deep intrinsic similarity between pleasant sensations. At the very least, these entities

[83] At this point, the reader may be inclined to object that it is impossible for a Martian to be disposed to seek out *the very same* kind of qualitative event that we have when we are being burnt. In any case, I ask that any worries of this sort be postponed until I can deal with them more thoroughly later, when I discuss the motivational system and phenomenology.

were posited because it seemed introspectively possible to come to dislike a core qualitatively identical sensation to the one we now like, and that the explanation for this was something phenomenological.[84] Whatever the hard to capture introspective datum was, it would be the reason for positing meta-sensations as theoretical entities. (And the same would go for any plausible alternative proposal about what constituted the difference between pleasant and unpleasant qualitative events.)

There seems to be the threat of a desire to have one's cake and eat it too here. On the one hand, there's the desire to use introspection to intuitively justify the positing of meta-sensations, but on the other there's the desire to pretend that we can't access qualia by introspection (as introspection is commonly understood), and that our only way of picking them out is via their connection to dispositions and behavior. Actually, this denial of introspection is a bit too quick on my part. It is obvious that we can become aware of pains without being aware of our dispositions as we become sophisticated cognitive subjects. To accommodate this evident datum, the account could allow a limited kind of introspection and introspectively based judgment (of the gerrymandered sorting kind discussed above) But whatever kind of introspection it was, it wouldn't be a strong enough kind of introspection to allow people to type pleasures together *on interesting intrinsic grounds*. Presumably, Robinson would be forced to such a limited introspection because he would have to deny, on this NIIC horn of the dilemma, that there was anything to introspect and type together on interesting intrinsic grounds where pleasures are concerned. (This is after all what is being claimed by the NIIC horn.) But then it is the same problem all over again—if there is nothing to type together on

[84] By 'core sensation' here, I mean the bare liked sensory content (e.g., the olive taste), abstracted away from the pleasant element.

interesting intrinsic grounds, then why bother positing meta-sensations? What explanatory work are they doing? Why not just have regular sensations, and simply say the sensations that we seek out or seek out the causes of are the pleasurable ones, and the ones we seek the end to are the unpleasurable ones? This approach would bring us very close to a return to behaviorism about pleasure and pain, and a reneging on the promise not to reduce pleasure to uncoerced SODs and seeking out behaviors. It is only by insisting on the *de jure* rigidity of concepts like PLEASURE that this behavioristic conclusion can be escaped, since this is the only way to avoid the consequence that pleasure and SODs will accompany one another in all possible worlds. I think that, at the very least, the arguments above do cast into doubt the acceptability of such an analysis.

A possible escape from this road beginning with worries about meta-sensations and ending in behaviorism about pleasure would be to insist, as suggested above, that the only initial motivation for positing meta-sensations was to explain how we could come to dislike a sensation we at one time liked, and vice-versa. So the reason for positing the meta-sensations would have nothing to do with any intrinsic commonalities between pleasures, but only with an ethereal, ineffable qualitative difference between an experience involving a core sensation that we like at one time, and an experience involving the same core sensation that we dislike at another.[85]

---

[85] I suppose that Robinson's arguments against the view that pleasure is a sensation could be seen as a rejection of interesting intrinsic commonality amongst pleasant experiences. In the relevant part of "What is it Like to Like?," as already discussed he does seem to suggest that we cannot isolate a common phenomenological element in all pleasurable experiences, at least not a readily describable one. (Take, for example, the claims about the relationship between the taste of olive and the visual experience of international Klein Blue discussed earlier.) But then again, in that section, he also seems to suggest that we cannot isolate any readily describable phenomenological element that distinguishes core sensations that we like from ones we don't that makes it the case that we like them. This a view he ultimately goes on to subtly reject as I understand it, or at least place heavy emphasis on the "ready describability." I'm not totally clear whether his later meta-sensation discussion (and, e.g., analogies with finding a face familiar) amounts to an "all things considered" revision of the NIIC claim, and a settling for a weaker claim to the

The main reason to reject this approach is that we run into the same introspection and gerrymandering worries as above. If all these meta-sensations are very different (and the total experiences of which they form a part very different) and the only reason for positing them is to explain subtle qualitative differences in liking and disliking the same core sensation at different times, we would expect introspective judgment that classifies qualitative states as pleasures to be labored, clumsy, and inefficient. But it is not. And moreover, we would have to deny the plausible *reductio* argument above in favor of a denial of the premise that pleasure is not a gerrymandered kind. I have already argued that this is not the most plausible move available.

As a result of all these concerns, trying to escape from the previous problems by insisting that we only posit meta-sensations to explain subtle phenomenological differences in liking the same core sensations at different times isn't a good idea.

So, to sum up, I think that the NIIC horn of the dilemma is vulnerable to serious worries about meta-sensations. In addition, the view of PLEASURE that most naturally leads to it is fundamentally flawed, and consequently a major motivation for accepting it is undermined.

Let's briefly examine the alternate horn of the dilemma, then—the one that the previous discussion suggests would likely be the more attractive one (and the horn that Robinson might actually be more sympathetic to, textual evidence notwithstanding). If the NIIC claim is false, then all pleasures do have some interesting intrinsic features in common. If there is some deep, intrinsic similarity among the pleasures, however, why could this not be the intrinsic positivity that we are seeking (or at least, why could

---

effect that there is a common intrinsic element in pleasures, but that this is something which is very hard to describe and misleading to call a "feeling" or "sensation."

intrinsic positivity not be part of that intrinsic similarity)? Granted, there are still hurdles to overcome. A couple of these hurdles (specifically, ones surrounding reactive dissociation cases and the phenomenological contributions of the limbic system) will have to wait until we discuss the relevance of scientific findings for the evolutionary argument. Others—worries surrounding the role of "cognitive" states in this whole process—will be dealt with and responses explored in the discussion of Robinson's two olive tasting cases coming up next. The issues arise most naturally in that context.

So in any event, hopefully I have shown that a potential space is available for intrinsic negativity and positivity, and none of the contrary considerations examined thus far have provided us with grounds to doubt that there is. In addition, I hope that I have avoided blatantly begging the question against Robinson by simply insisting that there obviously is something intrinsic that pleasures have in common, and that this is something we recognize introspectively.

Let me recap the findings of this subsection, before we continue on to the promised discussion of Robinson's two olive taste cases in an effort to show that something in the neighborhood of at least one of them is coherent after all. (In the process of this olive taste discussion, our ultimate goal will be to answer the objection to the evolutionary argument that claims that we do not have the concept of an IPP or INP, or at least no introspective reason to suppose that the concepts are actually satisfied by anything.) Recall that in this subsection, I examined two broad issues—whether Robinson's proposed analysis of PLEASURE as *de jure* rigid was tenable, and relatedly, whether it was plausible to suppose pleasures had anything interesting intrinsic in common. I gave three arguments against the *de jure* view, and contended that it was

doomed to fail for a variety of reasons, most notably its inability to account for the efficiency of introspective classification of pleasures. Along the way, I applied the lessons from these arguments (and provided an additional argument based on making sense of the motivations for positing meta-sensations) in an effort to show that pleasures have interesting intrinsic features in common.

The Problems with Robinson's Presentation of the Two Olive Tasting Cases

Now we are ready for the promised discussion of the two olive tasting cases. I hope to show that although the first olive case might be incoherent (or at least a scenario that no subject would ever be inclined to place himself in), something close to it probably is coherent. The reason is that Robinson oversimplifies the phenomenon of pleasure and phenomenal judgments about what we would intuitively describe as positive or negative qualia (contrary to the spirit of his own previous, subtle analysis). I also hope to show that the second case is probably coherent as well, on a plausible view of the nature of the phenomenology of pain and pleasure. Although I won't develop the view in its entirety that has the consequence that these cases are coherent, I will sketch out two slightly different models that track the potential directions the account could go in.

If something in the neighborhood of even one of these two cases winds up being coherent, and there really can be a mismatch between phenomenology and distal stimulus according to epiphenomenalism, then the evolutionary argument will be back up and running. This is especially so if the mismatch is due to the fact that intrinsically positive/intrinsically negative phenomenology can be accompanied by avoidance dispositions or SODs respectively on epiphenomenalism. The reason is again that,

granting that there is a strong connection between phenomenology and behavior according to physicalism, it will lead us not to expect IPP/SOD and INP/avoidance disposition matches (by the argument's lights) while epiphenomenalism will lead us to expect them no more or less than other matches. Thus, the fact that we actually do seem to observe IPP/SOD and INP/avoidance will count in favor of the non-epiphenomenalist theories.

So, let's begin by looking at Case A, the one where there is all manner of internal suffering that cannot be given overt expression. Jones is "inwardly railing" at the disgusting taste he's experiencing when he shoves olives into his mouth, but he continues to shove them into his mouth all the same. Recall that Robinson believes this case, as well as the one to follow, is incoherent, but that we are apt to slide between the two cases and illicitly convince ourselves that one (or both) of the cases really are coherent. (Of necessity, the discussion will also sometimes branch out into broader issues which will be dealt with in more detail subsequently.)

An initial and somewhat peripheral point to make is that it is not clear that appropriate verbal imagery (or any verbal imagery at all) would be essential to pain or pleasure. (Robinson doesn't claim anything to the contrary, of course, and is merely leaving open this possibility by considering Case A. But it is still worth the clarification, I think.) The analysis of pleasure Robinson proposes says roughly that a subject *a* is in pleasure iff *a* has a sensation *E* and *a* has *P* properly directed on *E*. *E* is caused by neural basis N(*E*), and *P* is a separate conscious occurrent, which we know is a pleasure meta-sensation from elsewhere.[86]

[86] Note that I continue my previous usage here by referring to the complex of sensation + metasensation as the "experience" or "qualitative event." Robinson's preferred terminology, because it is slightly different

Clearly, there is no explicit mention in these criteria of any subvocal verbal imagery, and it is implausible to suggest that the subvocal imagery is a component of either *E* or *P*. (The analyses Robinson provides both here and in "What it Is Like to Like?" certainly rule out that anything non-phenomenal could be essential to pleasure, though various non-phenomenal things could accompany pleasures in all nomologically possible worlds relative to the actual world. Nothing said is incompatible with that, but nothing said entails it either.)

In Case A, presumably the verbal imagery is supposed to constitute some kind of judgment or conscious desire (or both) about the taste. My real worry is that while inappropriate verbal imagery could be selected against (if there were a way for the verbal imagery to be inappropriate in the way suggested, which is an issue I will discuss shortly), it's not clear that it could be selected against if it were accompanied by other elements as part of a judgment. (A better way for me to put this might be that it would cease to be inappropriate if accompanied by these other elements, but again I will discuss this shortly.) For now, though, let me focus on the simple case, where judgment is just a matter of having the verbal imagery running through one's head—Jones has the phenomenal words "this is agonizing" in his conscious sphere, and that is what allegedly makes it the case that he judges the phenomenology negative or occurrently desires its absence.

A first thing that should be noted before proceeding further is that Robinson's attack on Case A's coherence is formulated in a bit of a confusing way, since it seems to suggest both that the subvocal sayings are cognitive, and that the causal efficacy of the

---

than mine, can potentially mislead here, since he speaks of the sensation and the meta-sensation as being separate conscious occurrents.

cognitive is undeniable, via its "products." As stated above when I first introduced Robinson's presentation of the case, we are understanding subvocal sayings to be a kind of verbal imagery, and after all, verbal imagery is just phenomenology in its own right, and phenomenology is causally inefficacious *ex hypothesi*.

To make matters worse, earlier in the paper Robinson suggests that thoughts are causally inefficacious. He says that "… it is somewhat mysterious how an occurrent belief or desire could causally contribute to behavior," and that, "it seems inefficient… to wait upon the formation of a thought before the organization of associated behavior begins. The more plausible picture is that the fit between our occurrent beliefs and desires, and our behavior, is a result of a process in which our brains… organize both our occurrent beliefs and desires and our behavior in parallel."[87] It certainly seems like verbal imagery constitutes the thought here, and aren't thoughts paradigmatic examples of cognitive episodes? And if the imagery doesn't constitute the thought here, what could constitute the thought if not the imagery, while still keeping the thought inefficacious? (Incidentally, regardless of the tenability of epiphenomenalism about verbal imagery in the face of evolutionary considerations, what Robinson says here about cognition is in independent *prima facie* tension with epiphenomenalism.)

A natural reply to the worries about inefficaciousness would be to claim that in spite of the inefficaciousness of the verbal imagery, its neural basis is causally efficacious. And since there is a nomic tie between this sort of phenomenology and the neural basis of it (i.e., the neural basis causes it in a lawful fashion), so long as the laws of nature are what they are, there will be a counterfactual dependence of behavior on phenomenology. Because natural selection selects for behavior, if this sort of

[87] Robinson (forthcoming b)

phenomenology accompanies appropriate behaviors, it will be selected for. End of mystery.

Whether or not this kind of reply is successful will depend on whether other mind-body theories lead us to expect a better match between phenomenology and distal stimulus, and hence make it striking that the correlations are what they in fact are if epiphenomenalism is true. (This is an issue that will be dealt with subsequently in this dissertation, and may be a serious problem for the evolutionary argument. For now, we are taking for granted that if we can make sense of the idea that pleasure is a kind of IPP and pain a kind of INP, then epiphenomenalism will have gained no dialectical advance against the evolutionary argument. And indeed, if this is the case, no matter which way other considerations lead us, epiphenomenalism will not have made dialectical progress as a result of the objection we are presently considering.)

Now, as I said above (when first discussing subvocal sayings), if the only accompaniment to the agonizing olive taste that could potentially strike us as out of place is the verbal imagery, then it will be very hard to come up with an argument for the coherence of Case A, at least in the striking form Robinson presents it in. If we stick to the claim that verbal imagery is all that there is to judging that the taste phenomenology is negative or occurrently desiring its absence or disliking it, then Robinson will be out of hot water and his contention that Case A is incoherent will be vindicated (at least if part of the understanding of Case A is that it really is in fact a negative judgment of the phenomenology). The reason is that verbal imagery is wholly constituted by phenomenal words going "through the head," so to speak. It would be hard to make the intuitive case that some variety of verbal imagery fails to match what we would expect to find being

caused by a particular stimulus, or resulting in a certain behavior. (This is true even if we grant the controversial assumption that there is a deeply satisfying account of why certain phenomenologies by their very nature cause certain kinds of behaviors on physicalism and interactionism.) After all, can we think of a plausible candidate for a mismatch scenario in this framework?

A suggestion would be something roughly like what Robinson says above—I hear the phenomenal words (e.g.) "This is agonizing," while I find myself uttering out loud "This is very pleasant" and picking up another olive. To get a better feel for some potential places where a mismatch could be thought to occur in this case and cases similar to them, let me sketch out a brief map of the various connections. I'll just list them:

(1) Between distal stimulus (taste bud stimulation) and neural basis of the taste sensation.

(2) Between taste bud stimulation and neural basis of the displeasure meta-sensation.

(3) Between the taste sensation neural basis and the taste sensation itself.

(4) Between the meta-sensation neural basis and the meta-sensation itself.

(5) Between these neural basis (or some combination thereof) and the neural basis of the verbal imagery.

(6) Between the neural basis of the verbal imagery and the verbal imagery.

(7) Between the neural basis of the verbal imagery and verbal behavior (overt speech).

(8) Between the neural basis of the taste sensation/the neural bases of the meta-sensation and motor behavior (the neural basis of the verbal imagery may be involved here as well).

(9) Between external auditory stimulus and neural basis of auditory experience (of hearing overt sounds).

(10) Between neural basis of auditory experience and auditory experience.[88]

All of these connections are going to be governed by various laws if epiphenomenalism is true, and in addition, there will be correlations between relata that figure in the laws, even if they are not causally related to one another in any direct way. (For instance, there will be correlations between phenomenological types and distal stimuli, and between phenomenological types and behavior. This is because, in the phenomenology/distal stimuli case, distal stimuli will interact with the physiological systems of the brain in a lawful fashion, and these will in turn cause the phenomenology in a lawful fashion. In the phenomenology/behavior case, the phenomenological types will be caused in a lawful way by physiological events in the brain, and the physiological events in the brain will in turn cause behavior in a lawful way.)

Returning to the olive scenario, let us examine several potential places where a mismatch could be thought to occur. (In order for such a mismatch to disconfirm epiphenomenalism, we would have to have reason to expect the laws we find that causally link mismatched things on physicalism, and we would have to have no reason—or at least less reason—to expect them on epiphenomenalism. Just to reiterate, the reason

[88] Nothing in this discussion relies on any very specific details of Robinson's meta-sensation account. An alternative for meta-sensations could easily be substituted in.

we are looking for a mismatch to begin with is that the general strategy of evolutionary arguments is to disconfirm epiphenomenalism by finding correlations that are unexpected on epiphenomenalism, but expected on the alternative—i.e., physicalism. These always involve mismatches, because there is supposed to be some special connection between phenomenology and behavior on physicalism, but not on epiphenomenalism.) The prime candidates for mismatch are (i) the verbal imagery with the motor behavior, (ii) the verbal imagery with the verbal behavior, and (iii) the total taste experience with the verbal imagery. (There are, of course, other candidates for mismatch in situations similar to this one—between distal stimulus and (non-verbal) phenomenology, for instance, or between (non-verbal) phenomenology and motor behavior—but these other kinds of mismatch are not at issue right now. For the moment, we are focusing on the possible mismatches that would involve verbal imagery.)

Motivating the idea that there is anything that could count as (i) is indeed very difficult. What is it intrinsically about *these* verbal images that makes them go poorly with grabbing olives and shoving them in one's mouth? After all, what if we had spoken a slightly different language, where 'agonizing' meant what 'pleasant' actually means? Then it would be natural, when entertaining verbal imagery in quasi-English in situations where our experiences were pleasant, to entertain phenomenal sounds like 'This is agonizing.'

(ii), on the other hand, is a different story. Although the case, as specified by Robinson, has the subject unable to say anything rather than saying something inappropriate, we can easily imagine a similar case that would suffer from this additional problem. It is much easier to motivate the claim that there could be a mismatch between

verbal imagery and overt verbal behavior than that there could be a mismatch between verbal imagery and motor behavior. But it is hard to motivate the claim that this mismatch could not be selected against if epiphenomenalism is true.[89] (And if the mismatch could be selected against on epiphenomenalism, we haven't yet found the right kind of mismatch for the evolutionary argument to stand a chance of succeeding. Again, the reason is that we are looking for a coherent mismatch—not found in actual human beings—that epiphenomenalism says would be as likely as any other correlation if epiphenomenalism were true, but which physicalism would claim is very unlikely compared to other correlations if it were true.)

The slightly modified scenario in question provides a very good example of such a potential mismatch—one utters things to oneself *sotto voce* that are not phonologically isomorphic to the things one utters out loud. In this case, one says to oneself 'this is agonizing', while one says out loud 'this is very pleasant'. The first utterance is broken up into 3 discrete phenomenal parts (THIS… IS… AGONIZING), whereas the fourth is broken up into 4 discrete units of sound (measured in whatever way these things are typically measured by sonic measuring devices). Also, the third word of the verbal imagery utterance is naturally divided into 4 parts which we refer to as "syllables," whereas the third word of the overt utterance is naturally divided into only 2 parts.

Of course, the mismatch is deeper than this. To illustrate, imagine I've just cut my hand badly on a sharp knife, and have the usual experience associated with such a cut. I have the English verbal imagery running through my head 'I'm in pain', but I find myself telling a passer-by 'this cut really hurts', even though no verbal imagery

[89] With Case A in its pure form (where the subject is unable to say anything), the obvious disadvantage is going to be inability to verbally communicate at all, which clearly could (and probably would) be a trait selected against in a given population.

syntactically similar to this statement came into my mind before I overtly uttered it. Plainly, there is a lack of syntactic or phonological isomorphism here as well, but we are much less likely to be struck by the inappropriateness as in the original example. Why? Intuitively, obviously it is because the two statements have related meanings in the cut case, but very unrelated (in fact, virtually opposite) meanings in the olive one. I won't try to speculate on what the synonymy relation amounts to, as doing so would threaten to take us far off course. For present (and very informal) purposes, it should suffice to suggest that there is some sentence bank or word bank (or some more sophisticated mechanism with the same basic function) embedded in our psychological architecture that allows us to interchange the words and sentences that we produce in ways that do not strike us as problematic. (I.e., they cause brain events to occur that result in similar functional results—similar behaviors, etc.) The cut case is an example of such an interchange, while the olive case is not.

So (ii) seems like a mismatch, but even if epiphenomenalism were true, we would still expect evolution to select against organisms that displayed such a mismatch just as well as evolution would select against them if the other theories were (even granting the controversial assumptions about physicalism's unbreakable tie between phenomenology and behavior). Consequently, the possibility of such a mismatch on epiphenomenalism does not provide evidence against the view.

The reason why we would expect evolution to select against organisms with this sort of mismatch even on epiphenomenalism is best illustrated by considering the following dilemma—either the neural basis of the verbal imagery is efficacious in some sort of behavior or it is not. (Presumably it is, but let's consider both options for the sake

of thoroughness.) If it is not efficacious, then the human brain would be constructed with a mechanism (likely a complicated one) for generating verbal imagery, but with this verbal imagery and its neural basis completely cut off from causing any sort of behavior, including verbal communication behavior. From an evolutionary standpoint, such a brain would be inefficient, since it would have to devote resources to sustaining mechanisms with no behavioral advantage. While it is surely possible that such a brain could evolve (since evolution does not always produce optimal organisms, and also it could be that the additional mechanism was somehow able to affect behavior in the past, but not any more), it is quite unlikely barring special additional evidence to the contrary. This unlikelihood is great enough to make it reasonable to dismiss the possibility. (Notice, of course, that nothing about this story is at all in tension with epiphenomenalism about qualitative events, any more or less than with any other theory.)

So much, then, for the inefficacious horn of the dilemma. Let's examine the efficacious one. If the neural basis of the verbal imagery is efficacious, then there will be a serious problem with communication. Though they are very different kinds of entities in very different mediums, both the verbal imagery and the overt speech are statements in the same language. Thus, the fact that one's overt utterances would have a completely different meaning from one's verbal imagery would likely lead to serious problems generally in coordinating actions with other organisms. For instance, if I form the verbal imagery "I'm feeling hungry" but constantly tell passers-by in overt speech that "I'm feeling full," chances are I will not be able to obtain food as efficiently as I otherwise would.

At this point, though, the issue arises of what makes it the case that my verbal imagery means something different from my overt verbal behavior. (And it is this difference that allegedly makes the mismatch to begin with.) After all, if both I and everyone else have roughly the same kind of brain and the same bridge laws operate for all of us, then what would make my verbal imagery 'hungry' (or 'agonizing') mean something different from my overt verbal utterances of 'full' (or 'pleasant')? Why would they not be more akin to the actual word pairs 'hungry' and 'starving', or 'agonizing' and 'excruciating'? So long as my verbal behavior causes the right kinds of neural states and behavior in my fellow organisms, no one will be any worse off. Won't this just be a case where we all speak a language where the verbal imagery 'I'm in agony' really means I'M IN PLEASURE, or perhaps the overt verbal behavior 'I'm in pleasure' means I'M IN AGONY?

This once again raises the issue of what is really involved in phenomenological judgments and "internal suffering." If the experience of the olive eater really does amount to nothing more than phenomenological raw feels (sensory and meta-sensory) + verbal imagery, then the olive example may collapse into Case B or something very close to it. In other words, it will be a case where verbal imagery and behavioral disposition are roughly the same as in the actual world, but where the meta-sensation has changed. (The one complication is that, unlike Case B, the verbal imagery won't literally be the same. In fact, the verbal imagery will have changed to strings of language we in English would think meant the opposite of what the old strings meant. The organism will say to itself the string of sounds 'this is agonizing', for example, rather than 'this is pleasurable'. But in her slightly different language, 'this is agonizing' will mean the

same thing as 'this is pleasurable' means in English.) This issue will come down to the answer to three important questions—(1) how does verbal imagery about phenomenology get its meaning—via connection to behavioral dispositions (or functional roles specified behaviorally), via its connection to the qualitative states it is supposedly about, or by some combination thereof? (2), does the phenomenology that itself constitutes the qualitative state of suffering (both the sensory and meta-sensory components) involve any representational elements? And (3), do these representational elements, or other qualitative representational elements aside from the verbal imagery, have any role in the judgments I make about my qualitative state, and if so, what role?[90]

Ultimately, verbal imagery's involvement in my judgment and (in addition to non-representational raw feel phenomenology) my internal suffering may wind up being only part of the puzzle. There may be other elements involved, elements whose representational content is somehow intrinsic to them, in the sense that it is not fixed by convention or by what behavioral dispositions it accompanies or behaviorally specified functional roles it fulfills. (In fact, these other elements may be all that is really involved in the judgments and the suffering.) In any case, all of these issues will be dealt with below.

---

[90] I won't get involved here in sorting out issues surrounding the extent to which my verbal imagery words like 'pleasure' and 'agony' have their meanings fixed by public criteria in the way that words of public languages presumably do. (Though the issues are certainly substantive, and worthy of further treatment elsewhere.) I also won't try to tackle the possibility that they are deferential in some way. Here, I will just assume that my verbal imagery language is the very same language as the one that gets overtly and publicly spoken, and that all the competent members of what we would intuitively identify as my linguistic community speak the same language. (I.e., different members don't speak phonologically and syntactically identical but semantically different languages, the differences deriving from their special private definitions of various words about mental things.) Of course, since presumably epiphenomenalism holds that all organisms of a species have similar brains, subject to the same physical/functional to phenomenal causal laws, there is no reason to suppose there will normally be great phenomenal diversity when they are exposed to the same kinds of stimuli, and so no reason to expect that their private definitions of words like 'pleasure', even if they did have them, would differ considerably from those of other community members.

Let us proceed on to an examination of (iii)—the claim that the mismatch would arise between the verbal imagery and the olive taste experience itself (where that experience includes both the sensory and meta-sensory phenomenological elements, but not the verbal imagery). In a way, though not itself very promising as a suggestion, (iii) indirectly helps us to focus our attention even more sharply on the issues just discussed.

On the face of it, there really is no potential mismatch of the sort (iii) specifies. Since, as we have already seen, verbal imagery is just a collection of phenomenal words, and words typically represent purely based on convention and the roles they happen to play in a linguistic scheme, pretty much any string of sounds could play the role that 'agonizing' or 'pleasurable' or 'hungry' or 'full' plays in English.

However, considering the possibility of (iii) does generate the intuitive reaction that something more is going on when a subject judges that she is in agony than merely having the verbal imagery 'this is agonizing' run through her mind. And moreover, the fact that 'this is agonizing' is English verbal imagery makes little difference—even if her verbal imagery were in some sort of mentalese, if this mentalese had meanings governed in the conventional and functional ways that English meanings are governed, the reaction would be similar. There is the intuitive feeling that, whatever it is that constitutes a subject's judgments about her qualitative events, it is something that does not represent in virtue of convention or behaviorally specified functional role (at least it doesn't present itself to her as representing in virtue of functional role.) *Even if she were a point of pure consciousness being deceived by Descartes' evil demon, the intuition goes, nothing would*

change about her ability to make the judgments she now does about the painfulness or pleasantness of her qualia, or have the conscious desires she now has about them.[91]

This is a very powerful intuition. If it is correct, there are at least a couple of different broad kinds of view that could accommodate it. One view is that the total qualitative event, the total experience (sensory component together with meta-sensory) itself, involves a representation of its own pleasantness or awfulness, and this is a major constituent in the subject's judgment that the state is pleasant or awful. In addition, of course, the state would normally also include an independent element of pleasantness or awfulness that would make the representation true. (David Chalmers, for instance, has suggested that sensations themselves are often constituents in the judgments we make about them.[92] I take this to mean that the sensations are the tokenings of the concepts we employ in the judgments, or at least a big part of those tokenings.)

Thus far, we haven't brought verbal imagery into the picture, but perhaps the verbal imagery 'pleasant' or 'awful' gets its meaning by standing in the appropriate relation to the pleasant/awful representational aspects of the phenomenology. In this case, I don't think the verbal imagery could be selected against on epiphenomenalism even if it was inappropriate, because the verbal imagery word would get its meaning for the organism from its connection to representations of goodness or badness coming directly from phenomenology, not from connections with SODs or the like. Thus, 'pleasant' would mean AGONIZING if it was related to negative representations, even

[91] Even if it is not metaphysically possible for someone to be in such a scenario (e.g., if people are necessarily constituted by their brains), there is some sense in which the scenario is epistemically possible which is strong enough to support the intuition that the relevant phenomenological concepts (PAIN, PLEASURE, etc.) would still function in introspection in the ways they actually do. This is all that is required for present purposes. And in any case, "brain in a vat" analogues probably do an adequate job capturing the relevant intuitions, and plainly are both epistemically and metaphysically possible in a recognizable sense.

[92] See Chalmers (2003)

though connected with SODs. This would, of course, make the word an item of private language that could only be gestured at suggestively by terms in public language, but behavioral dispositions, the sort of thing relevant to survival, could continue as before.

Or, the word 'pleasant' might get its meaning from the connection of the sensation to dispositions after all. It might have as its analysis 'sensation actually associated with SODs', or 'with the grounding of SODs'. But the public word (and the word of verbal imagery) would play no role in introspection or introspective judgment, since it would only point the way to an experience that had a way of representing itself, and this self-representation would be the real thing involved in introspection and introspectively based judgment.

Incidentally, I think this is the real lesson of the arguments against PLEASURE and PAIN being *de jure* rigid presented in the last subsection. The public words and even the verbal imagery words may be rigidified in the way Robinson suggests, but on the currently sketched picture of representation and phenomenology (which promises at least an intuitively satisfying account of introspection) these words do not express the concepts employed in introspection or introspective judgment, and the things that do are not rigidified in the same way.

Let's consider a slightly different view of phenomenology and representation that also accommodates the "evil demon" intuition about representation discussed above. Rather than representing itself as being awful or pleasant (i.e., having a component that represents the raw feel component as being awful or pleasant), the experience might simply *be* awful or pleasant, while some other phenomenological element (besides the verbal imagery) represents it as being such, and represents independently of convention

and functional role. This view would have the advantage of allowing pleasures and pains which are completely "raw feels," and don't just have some raw feel components—they lack any sort of representational content, any sort of content that is evaluable as true or false. Representation would then be an additional phenomenological event over and above the experience itself, potentially made true by the raw feel of the experience. (On the other view, pain and pleasure experiences always involve a representation of their own pleasantness or painfulness, a characteristic some might consider counter to the introspectively accessible nature of pleasures and pains. Incidentally, even if we do believe in these representational elements as parts of the pains and pleasures, we need not believe that the representation is always true—some experiences might include representational elements that are mistaken. There would be the issue of how introspection and introspective judgment could be so reliable, though, if they were accessing mistaken representations of the character of experiences. Perhaps the subject itself would have a way of correcting these misrepresentations when push came to shove in the process of judging, substituting a correct judgment for an incorrect one.) However, it would have the potential disadvantage of separating the representation from the pain experience (which might itself have distinct phenomenal elements comprising it, sensory and meta-sensory). The reason separating the representation from the pain experience might be a disadvantage is that presumably separating the representation from the experience would involve acknowledging that the representation has a separate neural substrate from the experience. Once this separate neural substrate is admitted, difficult questions arise regarding what role it plays in behavior, and whether organisms that have it might be selected against, either because it leads to behavior inconducive to survival or

because it is cut off from causing any behavior and thus is a worthless and cumbersome add-on mechanism to the brain.

In any case, I won't pretend to have cleared up all the problems with suggesting that pleasures and pains have IPPs and INPs respectively in the way outlined above. This is especially so for working out a picture of how these pleasures and pains might be consciously represented as having IPPs and INPs, in a way that introspection could make good use of. (After all, introspective judgments are a kind of representation in their own right, or at least essentially involve representation—they involve representation that a specified qualitative event or overall phenomenological state is occurring, and moreover that that qualitative event/phenomenological state is subsumed under a certain concept.)

And there are certainly other problems I haven't explicitly dealt with. One issue is that there is the risk that for the subject to be epistemologically justified or warranted in endorsing the representation, there might have to be some sort of mental act of "comparison" between representation and raw feel—at least in the case where the representational element is separate from the painful/pleasurable experience. If this were the case (assuming such a mental act of comparison was possible and could be made sense of on any mind-body theory), there would be a threat of the phenomenal elements exerting some sort of causation, which would violate the requirements of epiphenomenalism. (Issues like these will be dealt with in a later section, where the epistemological consequences of the global denial that qualitative events have causal powers is explored.)

Another issue is that the famous and extremely global "Kripkenstein" skeptical worry looms especially large in situations like this one.[93] What would make it the case that whatever we naively suppose is a conscious representation of the "painyness" of a qualitative event really was a representation? Why could it not just be a raw feel, with no truth evaluable content, in its own right?

Problems like the ones raised by Kripke's reading of Wittgenstein are very serious, and should not be minimized. At the same time, though, they are so global and insidious that they would undermine nearly all of what not only people in the street, but also philosophers, take for granted. Introspection and introspective judgments themselves, for example, would be out the window, since (as the discussion and proposed analysis of INTROSPECTION in the above subsection made clear hopefully) introspective judgment must employ concepts which do not have their reference fixed by anything publicly accessible.[94] (Basically because concepts that have their reference fixed by publicly accessible things, like behavior, cannot pick out phenomenal entities via the intrinsic features of those entities, which is what introspective judgment does. They can only pick them out via their accidental, relational connection to the various publicly accessible things being used to pick them out.) The kind of intrinsic representation many philosophers believe visual qualia to have, for example, would also have to be rejected.[95]

---

[93] See Kripke (1982).

[94] It is worth pointing out here that I have not always precisely distinguished between introspection and introspectively based judgment in the preceding remarks. As I am understanding it, pure introspection does not involve judgment—it is the mere "peering at" my phenomenal states (for lack of a better way to put it). Some philosophers have used the term 'introspection' in this way—e.g., see Gertler (2001)—while others have used it differently. Although I cannot do so in this work, I think it is well worth exploring the potential distinctions between introspection *simpliciter* and judgment about phenomenology based on introspection. For now, though, the level of precision being employed should be sufficient to clearly communicate and address the issues at hand.

[95] See, e.g., Horgan and Tienson (2002).

No longer could my visual qualia we would intuitively describe as of an oak tree in front of me be held to represent an oak tree in front of me, at least not by themselves.

But at the end of the day, there is still the "evil demon" intuition to contend with, and it is quite strong. We really do believe that even if we were in such a scenario, our pains would still feel negative, our pleasures would still feel positive, and moreover we could still introspect and recognize hurting and pleasantness—we could still represent qualitative events as being negative or positive, and represent them truly.[96]

So if we can work out one of the accounts of INP/IPP coupled with a view of introspection that acknowledges our ability to successfully represent INPs (e.g.) as INPs, it would open an avenue for something in the neighborhood of Case A that was recognizably coherent. In addition, the fact that human beings do not actually find themselves in situations like the one the case describes would disconfirm epiphenomenalism if we grant the controversial assumptions about physicalism (i.e., the unbreakable connections between certain phenomenal types and certain behavioral dispositions). The case would involve a person, when eating olives, having an experience that represents itself as being agonizing (or the person being phenomenologically constituted in such a way that some other, and presumably closely tied, phenomenological element represents it as being agonizing), and the experience actually *being* agonizing. This would allow the negativity of the experience and the judgment of its negativity to be a matter solely of phenomenology, independent of the behavioral dispositions associated with it. Thus, behavioral dispositions could remain the same,

[96] And also be justified in endorsing the representations, though addressing the justification of endorsing these representations would be a difficult subject in its own right. For the time being (at least until we deal much later with issues surrounding the causal powers of qualitative events), I am just assuming that the justification of our introspective judgments is unproblematic, aside from the brief discussion in connection with Case B below.

with no effect on the prospects of survival. But, in spite of this lack of effect on prospects for survival, there could be a noticeable difference in how the organism evaluates its own qualitative situation. (Noticeable, that is, from the organism's own perspective, not from that of a third person.) And the reason epiphenomenalism would be disconfirmed (in the actual world) is that this sort of mismatch between negative judgment and SOD is *not* found in actual cases, which is exactly what we would expect according to physicalism (granting the controversial assumptions) but not according to epiphenomenalism. (The reason I say only that this coherent scenario is in the neighborhood of Case A, and not that we have shown that Case A itself is coherent, is that recall Case A specified only that there was an alleged mismatch between verbal imagery and behavior. The newly sketched scenario is agnostic about whether there is a mismatch between verbal imagery and behavior, depending on how verbal imagery gets its meaning. If verbal imagery gets its meaning in the same way as public language does—via connections with behavioral dispositions—then no one would be tempted to say in verbal imagery the kinds of things Robinson supposes the person in Case A does. This is because 'agony' in verbal imagery would just mean something close to 'qualitative event actually associated in my psychology with activation of whatever the neural ground of extreme avoidance dispositions is', so it would never occur to me to describe myself as in 'agony' when I am seeking out olives, since this would be plainly false.[97]

[97] The obvious falsity of this claim assumes a number of things. First, that what qualitative events my neural events produce do not change willy-nilly over time, such that the same kind of qualitative event is not produced by different neural bases at different times—that would be contrary to any kind of plausible epiphenomenalist view, which posits stable bridge laws from physical/functional to qualitative. Second, that the neural grounding of the SOD has not changed—if it has, then there is the possibility that the old grounding produced what we intuitively describe as a "pleasure" when activated, but that the new one produces something else. And third, that when I actually am disposed to avoid something, the activated

If, on the other hand, verbal imagery gets its meaning from a connection with the (in this case) negative phenomenology or the conscious representation of the negativity of the phenomenology, then there will be a mismatch, but as I just argued it will not be a mismatch that can be selected against if epiphenomenalism is true.[98]

Now that we have examined Case A in some depth, it should be easier to consider Case B. Recall that Case B is much like Case A, except that there is no unusual verbal imagery. "The only difference from our case is that instead of N(P) causing P, it causes something else—let us say X. Maybe X is the same as what we would call displeasure, or maybe it is something else entirely." Recall also that Robinson goes on to claim that this possibility is incoherent as well, because (appealing to his *de jure* view of PLEASURE) "it is incoherent to suppose that X is the conscious occurrent that typically goes with activities we pursue without coercion, but is not pleasure." Using some of the results discussed both in the last subsection and in the examination of Case A above, I will argue that Case B is a coherent possibility as far as we can tell, and moreover one

---

neural grounding produces a different qualitative event than the activated neural grounding when I am disposed to seek it out. If not, then almost trivially any introspective judgments I make about being in pleasure or agony will always be true if I am in one of the two states, since pleasure and agony will amount to the same thing. We don't need any of these assumptions to establish that my judgment that I am in agony is unjustified, though—the only way I could be in agony while eating the olives on Robinson's *de jure* view is if one of the above assumptions are violated. But I would have no way of telling if one was violated, and it would seem unwise to suppose it was being violated without obtaining any evidence to that effect.

[98] Incidentally, in "What is it Like to Like?" Robinson actually considers and rejects the view that pleasure is representational, but the version of the view he rejects is very different from the one I have presented. That version is from Timothy Schroeder (see Shroeder (2001)), and holds that "an experience of a certain degree of pleasure is a perceptual representation of a certain quantity of positive change in one's net state of intrinsic desire satisfaction," where intrinsic desires are not to be understood as conscious events. None of what Robinson says in addressing that claim is applicable to our present discussion. In "What is it Like to Like?," Robinson is friendly to the idea that liking a sensation is an evaluation of it, though. He says frankly that he takes it that "liking a sensation is an evaluation of it," which seems to imply that the meta-sensation is in some sense an evaluation of the sensation it is about. He goes on to say that it is not an evaluation that can be true or false, though, no more than a seeming familiarity of a face can be true or false. It is only by going on to endorse (make a judgment) that I will like the sensation in the future, or that I have the seen the face before, that a mistake can be made. Thus whatever kind of evaluation the meta-sensation itself provides on Robinson's view cannot be pressed into service to perform all the desired representational tasks I outlined above as part of an account of INP and IPP.

whose absence from the actual world may disconfirm epiphenomenalism (by the evolutionary argument's lights), depending on how the details of the case are spelled out.

My first worry is that it isn't clear why case B threatens to be incoherent even *prima facie*. If having an unpleasant experience is just a matter of a certain kind of sensory + meta-sensory qualitative event occurring (as Robinson's account seems to suggest) and making a judgment about one's sensory state just a matter of having a certain kind of verbal imagery (as the original—and apparently intended—reading of Case A supposed), then the right thing to say here seems to be that the subject is having an unpleasant experience, but doesn't judge itself to be having one. As discussed in connection with Case A, Robinson's understanding of pleasantness and unpleasantness only involves appealing to sensory and meta-sensory qualia; it doesn't also involve appealing to verbal imagery.

It is a little difficult to discern, but I take it that the reason why he claims Case B is incoherent is that he is assuming that the creatures in Case B are actual humans, and thus (in keeping with his understanding of PAIN and PLEASURE concepts) it is an instance of the contingent *a priori* that whatever qualitative state humans are in when they have a SOD, it is a pleasure (supposing they are always in some qualitative state). In other words, on Robinson's view we can know that SODs are always accompanied by pleasures, just by reflecting on the concept PLEASURE. (To claim that actual humans are not in pleasure when they have SODs *is* incoherent on his theory of the concept of PLEASURE.) But the confusing thing is that he says there is a "difference from our case," which I take to imply that we are assuming that actual world human beings are as they are, and that this Case B subject lives in a world with a different bridge law from

physical/functional to phenomenal. Even assuming his theory of the concept is correct (and I have argued extensively that it is not), this seems like a perfectly coherent possibility, appropriately described as one where the alien subject is having an unpleasant experience, even though disposed to seek out the thing that gives him the unpleasant experience.

In spite of these difficulties, it is not too hard to see what Robinson is gesturing at here. He thinks that the neural basis of pleasure in actual humans could be nomologically connected to any of a number of different qualitative events, where the qualitative events are typed on intrinsic grounds. But no matter what qualitative event the neural basis is connected to, we have the same behavioral dispositions and the same verbal imagery (at least syntactically and phonologically the same). And if we have the same verbal imagery and the same dispositions, there is no way for us to notice these intrinsic differences, and hence no way that they could have anything to do with positivity or negativity (since positivity and negativity of qualitative events are palpably things that figure into our noticings and our introspective judgments if they are there at all). Even if we were wrong about this and there was intrinsic positivity or negativity, it still could not be used as part of the evolutionary argument, because it would be unnoticeable and hence not admissible as evidence.

My response to these suggestions shouldn't be too difficult to anticipate at this point. I don't accept the analysis of PAIN and PLEASURE this story presupposes because of the arguments in the previous subsection. Also, I take the possibility seriously that we could be aware of different intrinsic features of qualia and judge that these

differences were present even if our behavioral dispositions and verbal imagery (syntactically and phonologically specified) remained the same.

There is an important potential fallout of this discussion for my introspection argument in the previous subsection that should be dealt with, though. Recall that my dismissal of Robinson's view of PLEASURE and PAIN was largely motivated by that argument, and also by the argument that used some of the same introspection claims as part of a defense of the key premise that pleasure is not a gerrymandered kind. But in those earlier arguments, I assumed that introspective judgment is reliable, and moreover that we are justified in supposing it is. If Robinson's view of PLEASURE and PAIN and his apparent view about phenomenal judgment are correct, however, it is not clear how we will be justified in supposing that introspective judgment really is reliable. (His apparent view of phenomenal judgment is that it is just a matter of having certain verbal imagery running through consciousness.) The reason is that we would issue the same internal and external verbal behaviors no matter what the intrinsic character of our phenomenology, and we would be unable to notice anything wrong with our judgments if they really were wrong. (The only obvious way we could notice would be to claim we are in pain when we uncoercedly seek out the thing causing the pain, which would signal a violation of the contingent *a priori* truth that SODs never accompany pains. But an organism that manifested such behavior would certainly be subject to selection pressures.) The trouble is that we would be in roughly the same position with respect to our qualitative events as third parties—we would be picking them out via their connection with behaviors and behavioral dispositions. Consequently, we would be able to have no more confidence that we were correct than the third party—if the neural

grounding of our SODs changed over time, resulting in a transition from what we would intuitively think of as pleasurable experiences to painful ones, we could not notice. And if we plausibly insisted not just that concepts like PLEASURE were *de jure* rigid, but that their reference fixing occurred with an initial baptism (something like "the qualitative event actually caused by the neural basis of a SOD at t_0 when activated," rather than allowing qualitative events caused by a neural basis of a SOD at any time to count), the new qualitative events would not be pleasures at all.

If Robinson were to successfully defend a position like this, with a totally different conception of introspection than the one I presented in the previous subsection, it would spell systematic trouble for all my arguments. Robinson could simply reject my analytic claims about introspection in the first anti-*de jure* argument and escape its conclusion that way. And more than likely premise (2) in the gerrymandering argument—the one that claims that pleasure is not a gerrymandered kind—would be considerably less plausible as a result. It might lose enough plausibility to merit rejection, in which case Robinson would be able to hold on to his desired analysis of PLEASURE, which was the rejected premise (8). And of course premise (2) is very closely related to the "Nothing interesting intrinsic in common" (NIIC) horn of the dilemma for PLEASURE. Part of my strategy has been to first make the case that pleasures have something interesting intrinsic in common, and then go on to sketch a way for this something intrinsic to involve both a representational element that specifically represents the qualitative event itself as positive or negative and an element actually constituting that positivity or negativity (depending on whether it was a pleasure or a

pain). So if I can't even make the case that NIIC is false, my later arguments are going to have serious trouble getting off the ground.[99]

In response, I don't know that there's anything left to say on either side except to bite bullets. I'm not totally sure what I would say to someone who took seriously the skeptical possibility that our introspective judgments could be systematically mistaken, or were (contrary to the claims of the above argument) really reliable judgments, but judgments made on the basis of dispositional or outright behavioral facts.[100] (As discussed above, it seems obvious from what Robinson says in "What is it Like to Like?" that he is actually unfriendly to this latter alternative, in spite of its compatibility with what he says in "Evolution and Epiphenomenalism.") At least I can take some solace in the fact that the defender of a view like Robinson's who wanted to escape from my argument in this way would have to take seriously the epistemic possibility that our introspective judgments about pains and pleasures are systematically mistaken, or at least do not track the truth in a reliable way. As far as I'm concerned, that is a heavy burden to bear.

Now we have seen the problems with Robinson's presentation of Cases A and B, and can step back and appreciate the entire argument of the section. Throughout, I have endeavored to show that Robinson has failed to provide us with any reason to doubt that we have the concepts of an INP and an IPP, nor to suppose that the actual experiences we

---

[99] Though there are unrelated reasons to buy the later conclusions about intrinsic representation, such as the intuition that even if I turned out to be deceived by a Cartesian demon, I could still represent my qualitative events to myself as positive or negative. In a way, these intuitions and the anti-NIIC contention are mutually supporting—the ease with which I do the representing fits best in a picture where all the pleasures have something in common that constitutes the common representational content, and the seeming fact of common representational content helps to convince us that the qualitative events really do have something interesting intrinsic in common. As a result, it would be a serious setback for me to lose either of these general reasons.

[100] I will address the skeptical issue at greater length later in the chapter, but taking it on at this stage would threaten to disorganize the discussion.

intuitively suppose satisfy these concepts do not. I have contended that pleasures and pains respectively do have interesting intrinsic properties in common, and moreover that accounts of introspective judgment (and of concepts like PAIN and PLEASURE) that suggest that it works in some way other than by classifying experiences based on their interesting intrinsic properties are bound to be implausible. (Although I have not provided full accounts of INPs and IPPs, I hope that my outline of directions the accounts could proceed in is enough to convince the reader of the plausibility of the general proposal.) Consequently, I rest my case that the no IPP/INP concept objection to the evolutionary argument fails. Let us move on now to consider further objections.

III. The Empirical Objection to IPPs and INPs

The next objection grants that we have the concepts IPP and INP and introspective reason to suppose that those concepts are satisfied by actual qualitative events, but the findings of neuroscience give us grounds to doubt that our qualitative events really do have IPPs or INPs.

The primary neuroscientific findings that might lead someone to suppose this are ones having to do with "reactive dissociation" cases (sometimes called "reactive disassociation"). Reactive dissociations are experienced by some subjects under the influence of opium derivatives (especially morphine) or nitrous oxide, or who have received prefrontal lobotomies.[101] There are even reports of naturally occurring examples of the phenomenon, particularly in asymbolic patients.[102] Essentially, what apparently happens in these cases is that subjects experience very severe pain, but the pain doesn't

[101]See, for example, Dennett (1978).
[102]See Trigg (1970).

seem to bother them. In some sense, it doesn't hurt! Murat Aydede describes these oddities as follows, "[t]he phenomena seem to have a rather common phenomenology on the part of patients. In typical cases, they report quite sincerely that they have the pain as intensely as ever, but say that it does not bother them; they do not mind the intense pain they are experiencing, so to speak!"[103]

Subjects who are undergoing these dissociations often recognize that their insensitivity to pain is bizarre and abnormal, and realize that it is a point of interest for their examiners. Consequently, they sometimes feel the need to explain it to people around them. Apparently, they often insist that it is really pain they are feeling, and on occasion will even offer trumped up excuses for their insensitivity, such as that they are laborers who are used to painful injuries and so not inclined to pay attention to them.[104]

Reactive dissociation is so strange that some philosophers and scientists deny that the subjects' descriptions are accurate. Instead, they seek alternate ways to account for the data that don't involve such a gross violation of our introspective intuitions about pain. I will not entertain such challenges here—not because they are implausible (I won't make any judgment here about their plausibility), but because I don't believe the reactive dissociation evidence can be used as part of an objection to the evolutionary argument, even if the reality is what doctors and neuroscientists standardly take it to be. I will argue for that conclusion below.

Why, though, might reactive dissociation cases be thought relevant to evolutionary arguments against epiphenomenalism? One might contend that if these subjects really are in pain, and their pain doesn't hurt, that would show that pain isn't

---

[103]See Aydede (2000), p. 547.
[104]Trigg (1970).

intrinsically negative after all. It might seem like it to us introspectively, but the existence of these strange cases might provide grounds for thinking that there is a more complicated reality to pain phenomenology, hidden from introspection. These considerations might undermine our belief that there are reliable correlations between INPs and distal stimuli harmful to our prospects of survival, which would eliminate the key evidence in the evolutionary argument.

We can see a hint of this sort of approach in one of Robinson's earlier pieces on epiphenomenalism, an article for *the Stanford Encyclopedia of Philosophy*. There he says, "epiphenomenalists can meet [the evolutionary] argument… by supposing that both the pleasantness of pleasant feelings and the feelings themselves depend on neural causes (and analogously for painfulness and disliked qualities). So long as both types of neural events are efficacious in the production of behavior, their combination can be selected for, and thus the felicitous alignment of feelings with evaluation can be explained."[105] The idea here seems to be that once we appreciate the true, hidden nature of pain and pleasure (exposed by reactive dissociation cases and similar phenomena), we will see that it is not intrinsically negative, but rather has separable qualitative components that, once separated, no longer present an evidential challenge to epiphenomenalism. (These claims about separable components may strike the reader as more intuitive after the discussion of neuroscientific findings just below.)

Before continuing on to a discussion of the physiological basis for reactive dissociation (important for further fleshing out the relevant issues), it should be pointed out that reactive dissociation raises an important disanalogy between pain and pleasure, and perhaps between the entire phenomenological types we might naively consider INP

[105]Robinson (2003).

and IPP. The literature contains no reports of reactive dissociation cases for pleasure (as far as I know, at least)—there are no situations where subjects claim to be in pleasure, but don't enjoy it.[106] Consequently, any direct lessons of reactive dissociation cases relevant for our purposes will apply only to pain. Nevertheless, there are enough physiological similarities between the neural bases and etiological profiles of pain and pleasure that there are some lessons from the discussion that we can apply indirectly to the pleasure case.

So, we must now tackle in a little more detail the issue of what is occurring physiologically during reactive dissociation cases. In order to do so, it is important to get a basic feel for standard neurophysiological accounts of pain, in particular the "Gate Control Theory." (I won't get too deeply into the details here. Just the basic gist should suffice for our purposes.) The Gate Control Theory, which is very popular among neuroscientists and well confirmed in its broad outlines, posits the existence of a gating mechanism in the spinal cord that gathers information from peripheral nociceptors (i.e., nerves), generally outside the central nervous system. As its name suggests, this gating mechanism is essentially a modulating or regulating system that controls input from the peripheral nociceptors to neural structures deeper in the spinal cord. These deeper structures are the ones that eventually transmit information to the brain structures that form the basis of pain phenomenology.[107] Under normal circumstances, in order for the brain to interpret the transmitted signals as pain, the output of the gate "must reach or

[106]It is fairly uncontroversial that there are no known reactive dissociation cases for phenomenal types ordinarily considered positive. I am not sure if there are reactive dissociation cases for all types of phenomenology ordinarily considered negative, though. The only commonly reported reactive dissociations are for bodily pains.

[107] Aydede (2000).

exceed a certain critical level."[108] The output from the gate is also influenced by various competing excitatory and inhibitory factors.[109]

Most importantly for our purposes, stimuli from the peripheral nociceptors are routed from the gate to two distinct brain regions for processing. The first region is phylogenetically older—it includes the reticular formation, the limbic system, and the hypothalamus. Two pioneering researchers of the gating theory, Melzack and Wall, call these components together the "motivational-affective" system.[110] The second region is what they call the sensory-discriminative system. As Aydede explains, "[i]t involves the ventrobasal thalamus at which the noxious stimuli arrive through the spinothalamic and neospinothalamic projection systems and go directly to the somatosensory cortex, the basic *sensory* component of the system."[111] (There is also a monitoring and controlling link between the two systems, but the details of it aren't important for our purposes.)

Where pain experience is concerned, the primary role of the sensory-discriminative system seems to be in generating phenomenology associated with damage to various extremities of the body. The primary role of the motivational-affective system, on the other hand, is to generate phenomenology associated with motivation. As Aydede puts it, "[t]he point here… is that as an ongoing complex process, all these aspects and dimensions… are somehow fused with one another in our total pain experiences in a way that is often very difficult to distinguish introspectively, if not impossible. This is the

---

[108] Aydede (2000), p. 544.
[109] Though not relevant for our purposes, it is interesting to note that included among these factors are inputs from the brain, hinting—as common sense might suggest—that standing beliefs and desires, cultural factors, and past history can affect whether something is experienced as pain, and how intensely it is experienced as pain if it is.
[110] See, for instance, Melzack, Wall, et al. (1994).
[111] Aydede (2000), p. 545. Emphasis in original.

basis of the illusion that pain is essentially that singularly horrible, awful, abhorrent feeling."[112]

In reactive dissociation cases, what seems to be happening is that the input to the motivational-affective system (or the system itself) is being interfered with. Aydede claims that:

> … in reactive disassociation, the motivational-affective system somehow is not working properly as it is supposed to function, it is impeded, while the activity in the perceptual system remains intact, so that although the incoming signals from the periphery are processed and properly registered as pain along with its various characteristics in the perceptual system, they either do not reach the motivational system or do not produce their normal effects to activate it in the appropriate way.[113]

In fact, neuroscientists now know that most opium derivatives (including morphine) work by impacting the limbic system and various parts of the midbrain (central components of the motivational-affective system), while having little or no effect on the rest of the brain. And endorphins, the naturally occurring substances in the brain with chemical properties similar to opiates, are found mostly in these same regions.[114] With lobotomy, although there is no direct disruption of the limbic system or associated areas, the operation severs important connections between limbic structures and the frontal and prefrontal lobes, which normally provide a great deal of input to those structures. (Incidentally, severely depressed individuals often are unfettered by pain, and they often have limbic system impairments. Limbic impairments have also been implicated in at least some instances of congenital insensitivity to pain.)[115]

---

[112]Aydede (2000), p. 551.
[113]Aydede (2000), p. 548.
[114] Aydede (2000), p. 548.
[115] Aydede (2000), pp. 548-549.

As Aydede goes on to explain, "[a]pparently what makes pain experiences phenomenologically hurting, awful, or abhorrent, in other words, what makes them 'disliked', is the working of the affective/motivational system."[116] He also says that "[t]he phenomenon of reactive dissociation and the gate control theory of pain with its emphasis on parallel processing and interacting subsystems imply that the experience of pain does not consist, contrary to what introspectively it appears to be, of a simple and homogeneous qualitative feel. It is complex."[117] In reactive dissociation cases, the motivational-affective system is either impaired or starved for input, so the corresponding phenomenology has the sensory phenomenal pain aspect, without the usual affective phenomenology. (On Robinson's way of understanding these things employed in the last section, we might call this affective phenomenological component a 'meta-sensation'.) This sensory aspect of the pain phenomenology is affectively neutral—it is experienced as neither positive nor negative—it is only when added to the affective phenomenological element that the entire experience takes on the familiar negativity.

Let me offer a quick word about pleasure before going on to answer the objection to the evolutionary argument based on these reactive dissociation findings. Physiologically, pleasure seems to function somewhat differently, in ways that are perhaps not as well understood as analogous ones with pain. If one thing is well understood, however, it is that pleasurable experience is closely tied to stimulation of limbic structures. Aydede reports, for instance, that electrical stimulation of various limbic structures generates intense pleasurable experiences very reliably, even in serious

[116] Aydede (2000), p. 549.
[117] Aydede (2000), p. 550.

pathological depressives.[118] In some hard to describe sense, pleasure is much more a "reaction" to sensory information processing than it is part of that processing. (Aydede believes this may be the reason why it is possible to distinguish two simultaneous pains one is having, but not two simultaneous pleasures. To use Robinson's framework, the meta-sensation associated with pleasure has as its object a sensation that is more global and all-encompassing than the one associated with pain.) But although the process for pleasure is different in its details than for pain, it is still plain that in ordinary cases, total pleasure phenomenology is generated partly by neural events in the motivational system, and partly by neural events in other, more sensory oriented areas of the brain. For our purposes, this similarity is all that will be relevant.

Given my discussion of INPs and IPPs in the last section and my presentation of the physiological basis of pain in this section, my reply to the anti-evolutionary argument objection based on empirical considerations might be easy to anticipate at this point. A first thing to point out (that should be obvious from the previous discussion, and which is disputed by no one in the debate as far as I can tell) is that there is a phenomenological difference between a reactive dissociation subject who experiences pain and a normal person who experiences pain. Someone under the influence of morphine (e.g.), although perfectly able to classify his experience as painful, is not feeling exactly the same type of thing as the average person who sustains the same painful injury. After all, the average person's pain hurts, and the morphine patient's does not! There may be some phenomenal element that their experiences both share, but there is more to the one experience than there is to the other.

---

[118]Aydede (2000), p. 556. Aydede cites Buck (1976).

It is crucial to note, then, that the difference between me when I cut my hand, for instance, and my morphine using twin who is currently experiencing reactive dissociation is not solely a matter of our behavioral dispositions. (No one is disputing the claim that the limbic system, which is fully functioning in my case but not in my twin's, makes a contribution to phenomenology. This can be seen, for example, in the passage from Robinson's encyclopedia article above, when he speaks of the different neural causes of the pleasantness of the pleasant feeling and the feeling itself. Both the pleasantness and the feeling itself are obviously phenomenal, and according to epiphenomenalism the cause of the pleasantness is some neural event in the limbic system.)

But now let me assume the soundness of my earlier arguments whose aim was to establish that pains have something interesting intrinsic in common with one another (and *mutatis mutandis* for pleasures) and clear a place for the possibility that part of that intrinsic similarity involves a self-representing (and correctly representing) positivity or negativity. Even if reactive dissociation cases are as they seem, these arguments still hold for normal cases, where the motivational-affective and sensory-discriminative systems are working together. The motivational-affective system will make one contribution to the qualitative event, call it '*P*', and the sensory-discriminative system will make another, call it '*E*'. *E* and *P* will be bound together in a way which is undetectable (or at least very difficult to detect) in ordinary introspection, and if epiphenomenalism is true (implying that *E* and *P* are causally inert with respect to the physical), their neural bases will be working together to produce behavior. (It is highly plausible to suppose that the combination of neural bases is active in producing behavior, in particular when the sensory-discriminative system helps to cause activations in the

motivational-affective system, which the control system between the two allows when everything is functioning normally.)

But now the same basic problem arises on epiphenomenalism as arose before we knew about reactive dissociation and the complex phenomenology of pain. Even if one of the separable components is "affectively neutral" (neither positive nor negative), in normal cases the *entire experience* is not affectively neutral. There are actually two distinct possibilities, but both lead to this same conclusion.

We know that *E* is affectively neutral in and of itself (since this is what allows for reactive dissociation cases in the first place), so either *P* is not affectively neutral or it is. If *P* is not affectively neutral, then it is either positive or negative. When joined with a separate phenomenological component (*E*) that is affectively neutral, it creates an experience that is not affectively neutral. If *P* is affectively neutral (though various brain stimulation experiments suggest that it is not, since there can be objectless pleasures when the correct limbic structures are stimulated), then it nonetheless makes a non-affectively neutral experience when combined with *E*. No matter what, though, the experience is going to be affectively charged in a positive or negative way.

Depending on which of these options is correct, we can ask ourselves different questions. No matter what, though, the questions will be driving at the same point. If *P* is affectively charged, then we can ask ourselves why the motivational-affective system's neural basis of *P* (the one directly responsible for grounding behavioral dispositions presumably) causes *P* to be affectively charged in the way it is. If epiphenomenalism were true, why would the motivational system basis of an avoidance disposition be negatively charged, as opposed to positively charged or not charged at all? There seems

to be no reason to expect it to be on epiphenomenalism, since behavior would be the same regardless of the feel of the phenomenology, whether it was painful or pleasant or anything in between. But (granting the controversial connection between phenomenology and behavior) there seems to be great reason to expect it on physicalism. Consequently, physicalism is confirmed and the evolutionary argument survives.

If, on the other hand, *P* is not affectively charged, but just combines with *E* to make an affectively charged experience, then we can simply ask why the neural correlates of *E* and *P* together cause (e.g.) negative experiences when together they ground avoidance dispositions. The same issue arises—this is what we actually find, and there is no special reason to expect it according to epiphenomenalism, and reason to expect it according to the alternative.

We must acknowledge that epiphenomenalism *can* explain why organisms in normal circumstances don't suffer from reactive dissociation, and they can do this just about as well as the alternatives. As discussed above, reactive dissociation comes from an impairment of the motivational-affective system, or a starvation of input to it. If the motivational-affective system were impaired on a normal basis, the neural groundings of various behavioral dispositions would be destroyed. Presumably, no other groundings would take their place, and the dispositions themselves would fall by the wayside. But if this occurred, the effects could easily be devastating for the organism. If pain were experienced as just a sensational curiosity, with no motivational phenomenological component (at the neural level there would be sensory processing but no reaction in the motivational system and its hardwired connections to behavior), then organisms would fail to avoid stimuli that were threatening to their prospects of survival, and would thus

raise the risk of death substantially. (We need not look to hypothetical or counterfactual cases to make this point. Sadly, people who suffer from the rare neurological defect of congenital pain insensitivity actually do run into similar problems.)[119] Because the difficulties reactive dissociation subjects run into have clear physiological causes that all mind-body theories will roughly agree on, they will all give explanations of equal plausibility as to why reactive dissociation subjects are rare in ordinary circumstances. But when the theories try to generalize from that claim to a conclusion about pain and pleasure in all circumstances (including normal ones), the epiphenomenalist once again runs into the same old problems.

This concludes my discussion of the empirical objection to the evolutionary argument. Before considering further objections, however, I want to take a moment to deliver on a promise that was made in the previous section—to discuss further problems with accounts of introspection required to make sense of Robinson's *de jure* view of PLEASURE and PAIN in the light of the empirical considerations that have been presented.

Recall that in the last section I gave arguments that introspective judgment classifies qualitative events based on their intrinsic features, and moreover that pleasures and pains share interesting intrinsic features respectively. We are now in position to support those arguments with two additional arguments (which are not completely independent of the first arguments, but do strengthen them). The first is applicable to both PAIN and PLEASURE, while the second is mainly applicable to PAIN only (since it depends on the details of reactive dissociation cases).

---

[119] Interestingly, some pain researchers, such as Donald Price, do not call sensations that lack the characteristic affective component 'pain'. See Price (1999), p. 6.

The first argument is based on the fact that it is hard to believe that the motivational-affective system could be making similar contributions to behavior and a contribution to phenomenology in different pleasure scenarios, but that this contribution to phenomenology not be similar in the various cases (i.e., similar in interesting intrinsic respects). Although I take it as an open empirical question whether the exact (e.g.) limbic activation patterns are the same for all pleasures (or, more carefully, whether some core activation patterns are held in common), it seems plausible to suppose so, and the same goes *mutatis mutandis* for pains. And it does seem clear that the way the motivational-affective system grounds dispositions is fairly similar, given that the dispositions themselves are all fairly similar—they involve seeking out or avoiding specified stimuli, the only difference really being in intensity and urgency of seeking out or avoiding. But if the neural bases in the motivational-affective system were all fairly similar, isn't it plausible to suppose that the phenomenal components they give rise to would also be fairly similar, albeit hard to describe perhaps? In order for the phenomenal components to be dissimilar, something about the non-motivational phenomenological components they are combined with would have to substantially alter the way the motivational components present themselves to the conscious subject, or hide their true character. Although, if nothing else, reactive dissociation cases do teach us that phenomenology that appears unified may in fact involve discrete elements that have been subtly blended together, it does not provide precedent to suppose that features of the phenomenological components are misrepresented or hidden, just that the separability of the components themselves is not readily introspectible.[120] Consequently, I think we don't have good

[120]Other psychological oddities, such as perhaps blind-sight or experiments that involve flashing unattended-to visual qualia may have some relevance here, but I won't speculate further.

reason to suppose that if the motivational-affective neural basis is similar in different experiences, that the motivational qualitative component will not be similar also. Since I have contended that in fact it is plausible to suppose that there are common neural bases in the motivational-affective system of various pleasures and pains respectively, I conclude (at least provisionally, open to new information from neuroscience of course) that pleasures and pains share interesting intrinsic features with one another respectively.

The second argument is based directly on the reactive dissociation phenomena, and as I said above, is applicable only to PAIN introspective classification for that very reason. (And this argument is directed at introspection more specifically than the last.) It gives one more reason to doubt the presence of a dispositional record in the phenomenology used for introspective classification of pains, or in its neural basis. As we have seen, in reactive dissociation cases, patients are able to introspectively classify their pain phenomenology reliably. Although their pain doesn't bother them, they don't seem to have any difficulty recognizing that what they are experiencing is in some recognizable sense pain. As we have already discussed, it is plausible to suppose that these patients do not have full-blooded pain phenomenology, since they lack the motivational, meta-sensory contribution to phenomenology that accompanies normal pain. But they at least have the sensory-discriminative component of the pain phenomenology, and they are able to classify it as akin to the sensory component of full-blooded pain.

Now, by the very nature of reactive dissociation cases, these patients are not disposed to avoid the causes of their pain. In fact, they are indifferent to those causes. As a result, it is highly plausible to suppose that however they go about introspectively

classifying this qualitative state, their method does not rely on employing any information about behavioral dispositions, either consciously or at the subconscious neural level.

It could be objected at this point that really these subjects are subconsciously identifying a match between some aspect of the current phenomenology and some memory of a qualitative event that was associated with an avoidance disposition. (This would allow for successful introspective classification of pain without all pains needing to have anything interesting intrinsic in common with one another.) To use our earlier toy metaphor of compartments that house qualitative memories categorized by the dispositions they are associated with, the brain would be subconsciously detecting a match between the current experience and the sensory-discriminative phenomenological component of a previous qualitative event associated with an avoidance disposition. (Presumably that matching entry in the data bank would not have been gathered while the subject was under the influence of a drug that would be likely to produce reactive dissociation.)

Although I can't give a demonstrative argument against the claim that this is what allows the reactive dissociation subject to introspectively classify her pains, I am quite doubtful. For one thing, as discussed in an analogous situation in the previous section, there is still the problem of successfully classifying what we would intuitively consider pains that do not match any previous pains in their qualitative character. This problem would be likely to come up especially often in morphine related reactive dissociation scenarios. Since people under the influence of morphine are likely to be under the influence of that drug precisely because they have sustained an injury of a specially

unusual and serious kind (they have been operated on, have hurt an otherwise healthy body part in an accident, etc.), it would be surprising if they only experienced reactive dissociations (or were only able to correctly identify their phenomenology as non-hurting pain) in cases of pains fortuitously matching previously experienced pain in sensory-discriminative phenomenological character.

Plus, the old problem resurfaces of why introspection is so efficient. If the brain were trying to match the phenomenology of the new experience one-by-one against the voluminous variety of stored memories, one would expect the process to take far longer and be much clumsier than it is.

IV. Using Other People's Correlations As Evidence

Another potential problem with the evolutionary argument as given so far is that it relies a great deal on using information about other people's correlations (noxious stimuli with INP, beneficial stimuli with IPP) as evidence.[121] Presumably, this would include people living presently, and also ones living in the past. (Perhaps even information about ancestors to *homo sapiens* would be included.)

But, so the objection goes, what justifies me in using this data? After all, although I may be able to observe the sorts of distal stimuli other individuals are exposed to, and their behavior in response to them, I cannot literally observe the nature of their phenomenology. How do I know the neural states that give rise to their avoidance behavior, for example, don't also typically cause intrinsically positive phenomenology (where in me they cause INP)? In fact, if they did, everything would appear to me exactly the way it does, and so I would not be able to tell the difference. And on many

[121] I'm grateful to Frank Arntzenius for pointing out the need to answer an objection based on this issue.

views about evidence, it would be highly improper for me to count something as evidence when a relevantly different state of affairs (from the one I am taking to be the case and using as evidence) would be indiscernible from my perspective if it in fact obtained. But if I can't use these correlations as evidence, it seems the argument is in trouble, since the vast majority of the empirical data it depended on will be out the window.

Fortunately, I think there are plausible responses to this objection. The first (and what I take to be the preferred) response is to appeal to simplicity considerations, in much the way many philosophers do to counter skepticism about other minds generally. It is apparent that, for the most part, people behave in similar ways when exposed to similar stimuli, at least in the more dramatic cases. When any two people are burned with a hot iron, for example, typically they both jump away with similar urgency. Moreover, enough work has been done by neurobiologists over the past century and a half that we can be reasonably confident that when two typical people are exposed to similar noxious stimuli (especially of dramatic sorts like hot irons, stove tops, or dagger wounds), the neural events occurring in their brains and nervous systems are normally very much alike. But it would be exceedingly odd, then, if in spite of having similar neural bases, these similar neural bases gave rise to entirely different kinds of phenomenology in different individuals. In order for this to be the case, the laws of nature (in particular, the bridge laws from physical/functional to phenomenal) would have to be gerrymandered in strange ways—spatiotemporal location and/or which individual was involved would be an important factor in determining what phenomenology was produced. (If, for example,

the neural basis occurred in one room, it would give rise to pleasure, but if it occurred in another, it would give rise to pain.)[122]

If psychology posited such laws, the sort of gerrymandering involved would be unprecedented in the sciences. There doesn't seem to be any reason to suppose that spatiotemporal location would be a parameter in bridge laws from physical/functional to phenomenal—positing such laws is not required to accommodate any known psychological data—and laws with these additional constraints are less simple than alternatives without them.[123] Since it is standard scientific and general inductive practice to prefer simpler hypotheses to more complicated ones *ceteris paribus* (assign them higher prior probabilities, if you like), and the more complicated hypotheses in this case do no better job accounting for the data, there are no grounds to prefer them.[124] (How could they do a better job of accounting for the data, after all? We are exploring epiphenomenalist hypotheses here, and according to epiphenomenalism, the qualitative events being produced by the posited laws are causally inert with respect to the physical, and so cannot causally affect the things capable of interacting with the observations that provide us with or constitute our evidence.)

But then it is a straightforward step to complete the response. Since I am aware of what qualitative events are occurring in my own case when I'm exposed to stimuli that threaten my survival (or so we are assuming for the moment), and justifiably confident

---

[122] Recall that epiphenomenalism (any sort of mainstream version of the view, anyway) requires that phenomenology be caused by physical/functional events in the brain, and so presumably that this causal process be governed by laws of nature.

[123] I will not attempt to give an account of what makes one hypothesis simpler than another, as this project is mired in well-known difficulties that can't be dealt with here. I will simply rely on an intuitive characterization.

[124] I have spoken a bit too quickly on the simplicity issue here, since it is standard practice to prefer a more complicated hypothesis *ceteris paribus* in some isolated cases—like when it coheres better with a broader theory or our background beliefs about the nature of the world. But it is clear in this case that the gerrymandered bridge laws enjoy neither of these advantages.

that my own neural states roughly match the neural states of other typical people placed in similar circumstances, I can be justifiably confident that those other people's qualitative events are similar to mine in the relevant respects (i.e., intrinsically negative). This is because, in keeping with standard inductive practice, I can assume that the bridge laws from neural states to phenomenology are not gerrymandered in a way that would give rise to different phenomenology in the different cases. (And it is not much of a jump to suggest that, just as I can be confident of my relevant similarity with present persons, I can be justifiably confident of my relevant similarity with past ones as well, since there is no good evidence for thinking that those past persons are physiological different from present ones—again, in any relevant respects—and to suppose otherwise without evidence would itself be an unsimple hypothesis.)

The objection can be pressed further, of course, but only by raising skeptical worries of a more global character—worries about induction, or about our *ceteris paribus* preference for simpler hypotheses over more complicated ones. Although I cannot refute these more insidious skeptical problems (just as I could not refute the Kripkenstein worries raised in a previous section), they are general issues that plague human inquiry, and are not at all specific to the current debate. Not that they are automatically unreasonable as a result—they may well be very reasonable—but appealing to them in the present case feels a bit like objecting to the historical claim that Oswald shot Kennedy on the grounds that we have no reason to believe in an external world. The objection seems out of context, and calls into question assumptions that are being treated as axiomatic by both sides of the debate.

For those unpersuaded by the specific simplicity response in this instance (but who are not so globally skeptical that they reject induction, or simplicity responses in general, or the existence of the external world), there is a second response that may strike them as more attractive. (It is more of a "damage control" response than a straight reply to the objection.) In the next chapter, when I discuss alternative forms of the evolutionary argument and variations on the same theme, I will argue that there is a version of the argument which dispenses with evolutionary evidence altogether. The only evidence it uses is that people *now* by and large manage to survive for long periods of time in spite of many noxious stimuli and threats to their survival (and that these individuals have the correlations between survival threatening stimuli and INP that we common-sensically suppose—and the same goes *mutatis mutandis* for survival prospect enhancing stimuli and IPP). Although I can't get into the details here, in that chapter I argue that this version of the argument, with its thinner evidence, is about as good as the original evolutionary argument; the loss of evolutionary evidence does little to change the argument's force. If I am correct about this, however, then it seems an even more restricted argument could be formulated to bypass the present worry (albeit with a less general conclusion than the original argument—hence the "damage control" aspect).

Imagine only using one's own first-person correlations between phenomenology and distal stimuli as evidence, plus the further evidence that I (i.e., the person running the argument) have gone at least two or three decades without perishing, despite encounters with numerous life-threatening stimuli. (I ask those readers not persuaded that the loss of evolutionary evidence is fairly benign to set aside their worries until they can be discussed in the next chapter.) If the basic dialectical move of the original argument is

sound—i.e., that physicalism leads us to expect the very correlations we find (because if phenomenology were different, different and less fit behaviors would result), while epiphenomenalism does not—then I should at least be able to arrive at the conclusion that epiphenomenalism is disconfirmed vis-à-vis physicalism as a theory about *my own first-person* case. After all, if the INP/survival threatening stimuli correlations hold in my case, and physicalism entails that any deviation from these correlations would have led to different, less fit behaviors, and in the two or three decades I have managed to survive those less fit behaviors would have manifested themselves in destructive ways, then physicalism does a good job of leading us to expect the evidence. (It makes that evidence likely if the hypothesis is true.) If epiphenomenalism, on the other hand, entails that any of a substantial number of phenomenological variations (many of which were not intrinsically negative) would have resulted in the same behavior, and so would not have affected my survival prospects, then it does not lead us to expect the evidence. (It does not make the evidence likely if the hypothesis is true.)

Now although I cannot obtain much support for the anti-epiphenomenalist conclusion generally (as a mind-body theory for the human species) without appealing to either simplicity considerations or admitting evidence about other people's phenomenological/external stimulus correlations straight out (moves we are disallowing), this does not ruin the support it gets in the first-person case. (Again, it bears repeating that we are assuming that the main dialectical move of the argument is correct.) And even though this conclusion isn't very general, it is nevertheless interesting that I, and any person relevantly like me, can run an argument like this and disconfirm epiphenomenalism in the first person case. If, after scrutiny, the basic move of the

argument holds up, then even those individuals hostile to simplicity considerations and the like can still salvage a noteworthy result.

I hope the reader is satisfied, then, that this objection can be overcome in one of two general ways. It can be overcome directly by appealing to simplicity considerations and the like, or it can be overcome indirectly by using first-person evidence.

V. Introspective Judgment and Epiphenomenalism

The discussion about use of first-person evidence may have sparked a question in some readers' minds—if epiphenomenalism is true (a possibility that is certainly being left open), what would justify me in even using observations about my own phenomenology as evidence?[125] After all, epiphenomenalism seems to entail that phenomenology isn't causally efficacious in producing judgments about phenomenology. (Epiphenomenalism, as we've understood it here, trivially entails that qualitative events are causally inert with respect to physical events—this is part of the definition of 'epiphenomenalism'. It is only with additional assumptions, albeit highly plausible ones, that it further entails the inertness with respect to qualitative events.)[126] But if that is the

[125] In keeping with earlier discussion, the kinds of phenomenal judgments I will treat as relevant here are first-person ones that employ concepts different from the ones we use to pick out phenomenal entities in public discussion. (This tracks the distinction in Chalmers (2003) between "pure" and "impure" phenomenal concepts. An example of an impure phenomenal concept is the one expressed by the English word 'red', as that word applies to a phenomenal quality, rather than a property of external objects. The term has its reference fixed *de jure* rigidly, via something like the following description—'the visual phenomenal quality paradigmatic objects—fire engines, stop signs, etc.—cause in most people in lighting conditions of such and such sorts'.) I also set aside phenomenal judgments arrived at in ways obviously crucially different from introspection, regardless of the sorts of concepts employed. For example, a neuroscientist who is currently in pain, but forms a judgment about the pain not by introspection, but by making an inference from some CT scan readings he is currently examining is not making a phenomenal judgment of the sort that will concern us, even if the PAIN concept he employs is appropriately divorced from public reference-fixing criteria.

[126] See the earlier discussion on epiphenomenalism and its motivations for a discussion of this point.

case, it might be natural to suggest that judgments about phenomenology are themselves unjustified, since the phenomenology played no role in causing them.[127]

We have now come to a turning point in our discussion. Although some of the objections considered thus far have required lengthy refutations, the difficulties for the argument have not been great in principle (except perhaps for objection (D), but the worries encountered there weren't well developed enough to give an ultimate evaluation). From now on, though, we will be looking at objections which threaten the core of the argument much more urgently (and perhaps successfully).

As I alluded to briefly earlier, making sense of introspection and judgment about phenomenology on epiphenomenalism is potentially a very troublesome endeavor. To do the matter any justice, at least a book-length treatment would be required, and background work on the nature of introspection, phenomenal judgment, and even experience more broadly would have to be carried out. Here is unfortunately not the place for such comprehensive discussion. For now, my intent will be to make it clear that the sheer fact that phenomenal to physical causation is ruled out on epiphenomenalism, and phenomenal to phenomenal causation rendered very difficult to countenance, is not enough basis to confidently claim that judgments about phenomenology are unjustified if epiphenomenalism is true, or that they cannot be used as evidence. Actually, strictly speaking, my conclusion will be even more modest—that there is no good reason to suppose that various leading paradigmatic theories of epistemic justification (to be introduced and discussed shortly) come down differently on the justification of phenomenal judgments given that epiphenomenalism is true rather than physicalism. In

[127] See McLaughlin (2006) for a brief statement of this worry (though not one specifically directed at the evolutionary argument).

other words, it is reasonable to suppose, at least on a cursory inspection, that these theories of justification would conclude that relevantly similar phenomenal judgments on epiphenomenalism and physicalism respectively are either both justified or both unjustified. While this does not ultimately constitute a complete and fully satisfying response to the objection, I hope that it is at least enough to allay the *prima facie* concerns that the reader might have about introspection and introspectively based phenomenal judgment on epiphenomenalism.

Before beginning, I should make several brief comments to clear up some potential confusions and objections. First, a quick terminological point—I will continue to use the word 'judgment' as I have throughout, to pick out purely conscious, occurrent beliefs. On occasion, this may make for some slightly awkward phrasing—e.g., 'justified judgment'—where readers will be accustomed to seeing words like 'belief' instead of 'judgment'. I have accepted this awkwardness knowingly, supposing it worth the tradeoff for terminological consistency.

Second, throughout the discussion, I will assume that the question of whether a phenomenal judgment can be used as evidence is equivalent to the question of whether it is epistemically justified. This is a non-trivial (and indeed, non-obvious) assumption for a number of reasons. For one thing, it is typically thought that what an agent is entitled to use as evidence is intimately tied to what an agent can believe without being epistemically reckless or irresponsible, where recklessness and irresponsibility are ultimately cashed out in a way that places much emphasis on the things the agent has direct cognitive access to. In this day in age, on the other hand, epistemic justification is sometimes construed as having little to do with satisfying epistemic obligations like

responsible belief and weighing of reasons for and against. (This is true particularly for the more radical externalist theories of justification—i..e., roughly those theories that claim justification has little to do with an agent basing judgments on reasons or having reasons cognitively accessible to him.)[128] Although some theories of epistemic justification still tie justification in with the satisfaction of specific internally accessible prescriptive constraints, this is no longer taken for granted.[129] Thus, to take a fairly dramatic example, although some epistemologists might be willing to claim that a reliable clairvoyant has justified judgments about happenings in a part of the world she has never been to and never received information about, the same epistemologists might be very squeamish about acknowledging that the same clairvoyant was entitled to use her judgments as evidence.

Another potential problem with equating justification with being entitled to use something as evidence is that many people have the idea that evidence is something which the agent must be "certain" of (whatever exactly certainty amounts to, a very thorny issue).[130] A judgment, on the other hand, can be justified without the agent having maximal confidence or impeccable reasons for it. (Indeed, on some externalist theories, without having any reasons at all—though some degree of confidence is always required presumably, otherwise there would not be a judgment to assess in the first place.) The main reason for this insistence on certainty is that it seems to be required in order to

---

[128] Of course, this is not to deny that even externalists typically appropriate some normative role for the attainment of justification—insofar as such theorists think knowledge is something good and worth attaining, and insofar as they believe justification is necessary for knowledge (and sufficient when combined with truth, plus whatever is required to alleviate Gettier problems), they will clearly think that justification is at least an instrumental good. It allows for the attainment of knowledge. My only point here is that this will often be a kind of normative achievement that is not largely under the agent's control.

[129] For an interesting discussion of epistemic deontology and its historical role in shaping theories of justification, see Plantinga (1993).

[130] Some might say "appropriately certain" rather than "certain" simpliciter, the point being that mere maximal confidence is not enough to ensure certainty in the relevant sense.

avoid various skeptical challenges, which would be a serious obstacle to satisfying the demands of epistemic responsibility. (And as discussed above, satisfying requirements of epistemic responsibility are typically thought to be intimately connected with satisfying requirements for using a judgment as evidence.)[131]

There are familiar ways to deal with the certainty issue, however, without yielding to the skeptical challenges. If one's picture of the transmission of evidential support requires that beliefs be appropriately supported by foundational evidence, and one's picture of appropriate support is Bayesian, then Jeffrey Conditionalization can accommodate evidence which is itself uncertain. Although accounts of Jeffrey Conditionalization are plagued by a number of difficulties, if supporters could iron out these difficulties (as they have begun to do in recent years, at least in outline) this disanalogy between requirements for justification and for counting as evidence would be eliminated.

In any case, I won't pretend that I can provide any sort of knockdown argument for equating the entitlement to use a judgment as evidence with its epistemic justification. For present purposes, as I said above I'm just going to assume that there is this link between evidence and justification, primarily because there is a much more substantial literature on epistemic justification than there is on the nature of evidence, which makes it easier to address various paradigmatic dialectical options. For those unwilling to permit this assumption, it should be noted that theories of evidence requirements are typically much more closely aligned with internalist theories of requirements for justification (since, as noted above, these internalist theories usually emphasize

[131] Incidentally, a possible exception to this is the account of evidence provided in the work of Timothy Williamson, especially Williamson (2000). Williamson's account is novel, sophisticated, and intricate, so I will have to set it aside for the time being due to lack of space.

fulfillment of responsibility requirements or the like). Consequently, readers of this ilk can simply ignore discussion of the more externalist theories as irrelevant, since equating justification with entitlement to use as evidence (or claiming that evidence and justified belief are coextensive, if you like) only broadens the field of dialectical options—options for the epiphenomenalist potentially to get into trouble on. Since the point of the section is to argue (at least preliminarily) that phenomenal judgments based on introspection would not have their evidential status threatened by the truth of epiphenomenalism (either because they already lacked such status, or because epiphenomenalism did nothing to take it away), narrowing the field of possibilities can only strengthen the overall conclusion of the section. (The remaining members of the field will be the more difficult alternatives for epiphenomenalism to deal with anyway.)

A third preliminary issue worth remarking on is that, in addition to questions about the evidential merit of introspective judgments on epiphenomenalism, there are overall questions about the evidential merit of judgments about the relevant (e.g.) INP/ survival threatening stimuli correlations, which are ultimately the things that get used as evidence in the anti-epiphenomenalism argument (the first person version, anyway). Answering this overall question requires addressing issues about the evidential merit of judgments about the external world and their relationship to phenomenology if epiphenomenalism is true, in addition to issues about introspectively based phenomenal judgment. I will not address these further issues here, preferring to restrict my attention to ones surrounding introspection. Aside from the obvious reason—spatial constraints—there is an additional consideration that prompts me to avoid the further matter. Basically, if we can make sense of introspectively based phenomenal judgment, there will

be little interesting difficulty in principle to making sense of the external world judgments and the correlations between them and phenomenology. If an externalist theory is correct, then presumably (supposing there really is an external world) if the theory makes phenomenological judgments come out as justified on epiphenomenalism, it will also make judgments about the external world and correlations between the external world and phenomenology come out as justified also.[132] (This is because it is reasonable to suppose the same kinds of reliable mechanisms or counterfactual dependencies will hold—more on the details of the various theories later.) If, on the other hand, an internalist theory is correct, then if it yields the conclusion that introspective phenomenal judgments are justified, any beliefs about the external world and correlations between events in it and phenomenology will proceed via familiar sorts of inductive and abductive inference from the phenomenal judgments. (E.g., inferring that phenomenology of this particular sort is most likely to be caused by a stable world of physical objects, etc.) As I discussed in the previous section (and which is news to no one), there are difficult skeptical problems surrounding these kinds of inferences. But these skeptical problems have nothing to do with epiphenomenalism specifically. (There is one possible exception—perhaps the issue of making sense of inference on epiphenomenalism is a difficulty, given that causation between judgments cannot be countenanced, and it might be plausibly contended that in order for a judgment B to be inferred from a judgment A, B must be caused by A in the right sort of way. There may be good responses available to the epiphenomenalist here, but in any case I have to set

[132] I am assuming that the correct theory of justification/evidence is not weirdly gerrymandered, providing (e.g.) different kinds of requirements for different kinds of judgments.

the issue aside, as it is systematic enough to be well beyond the pale of the present discussion, which is admittedly preliminary.)

While we are on the topic of radical skepticism, one final preliminary comment is in order. In the previous section, part of my strategy for bypassing the objection against using other people's distal stimulus/phenomenology correlations as evidence was that it involved appealing to a radical skepticism that was out of place in the present context. But, skepticism about the evidential merit of phenomenological judgments is itself a very radical skepticism, plausibly much more radical and fundamental than skepticism about the existence or nature of other minds. Why not just get around the phenomenological judgment objection, then, in the same way as the objection about using the correlations in others—by denying that the kind of skeptical worry raised is appropriate in the context of the debate? The reason is that, in the last section, the kinds of skeptical worries raised had little to do specifically with epiphenomenalism; they were just general skeptical problems (like problems of induction and abduction) that universally plague attempts to make inferences to the best explanation. But the skeptical problems surrounding introspection on epiphenomenalism are clearly much more centrally motivated by specific concerns about epiphenomenalism. Consequently, it is much easier to defend the intuition that they should be addressed by an argument that takes epiphenomenalism seriously as a dialectical option in the mind-body controversy.

Now that we have prevented some potential preliminary confusions and objections, let me continue on to the main task. Examining whether epiphenomenalism affects the prospects of a specific kind of judgment's epistemic justification naturally requires articulating a theory of epistemic justification and evaluating the judgment with

respect to it (assuming alternatively physicalism and epiphenomenalism). And it should come as no surprise that there is great diversity not only in the sorts of theories that have been advocated, but even in views about the appropriate methodology for picking a theory. Some philosophers have thought that we should fit our theory to a set of paradigmatic particular cases of justified judgments. Others have thought that we should obtain principles of justification first (gotten perhaps from a theory of epistemic normativity, or intuited directly), and consider particular cases only after questions about the general principles have been settled. Others have thought the best method is a combination of the above approaches, fitting some rules to particular instances of justified judgment whose justification we are unwilling to give up, and ruling out other particular cases from justification because they conflict with abstract principles we refuse to dispense with. (This is often called the "reflective equilibrium" method.) The debates have become further layered in complexity, because often various competing theories have had subtly different notions of justification that they are trying to analyze—some have wanted to give an account of justification in a rigorous philosophical context, others to give an account of justification as it ties in with ordinary people's attributions of it in everyday situations, and still others to articulate a revisionary understanding, as part of an effort to revamp ordinary practice in a direction friendly to "naturalism" (whatever exactly this amounts to).[133]

In any case, I certainly won't try to settle any questions here about the correctness of any of the competing theories, preferring instead to select several paradigmatic

[133] Not surprisingly, epistemologists trying to analyze the more rigorous kinds of justification have leaned in internalist directions, the everyday in fairly moderate ones, and the revisionary in externalist ones. (Though, as with most broad generalizations, this rough and ready claim is subject to numerous qualifications and exceptions.)

theories of justification from among the myriad of options along the continuum (from internalist to externalist), and investigate whether introducing epiphenomenalism is of any consequence on any of the respective theories. While this approach is certainly not exhaustive, and cannot serve as the basis for any definitive conclusions, it should be sufficient for present purposes, which as I indicated are not intended to be anything more than preliminary.

I will discuss four separate theories, one markedly internalist, a second fairly moderate, and two externalist (the two externalist ones serving as exemplars for two prominently different varieties of externalist option). The internalist view will be a version of strong internalist foundationalism, the moderate view a version of a defeater theory, and the two externalist views process reliabilism and a counterfactual dependence theory inspired by Robert's Nozick well known account.

I will begin with the internalist option.[134] Strong internalist theories are notoriously difficult to pin down precisely, and just formulating internalism coherently could take the better part of a dissertation. For that reason, I won't try to take on the mammoth project of formulating the view in detail. The version I sketch here may not be perfectly precise, but it should allow us to concentrate on the salient issues.

So, for now, let us understand strong internalist foundationalism as follows:

---

[134] A view along these rough lines is defended in BonJour (1999). A similarly strict internalism is defended in BonJour (1985), though there BonJour advocates an internalist coherentism rather than a foundationalism. The reason I do not consider an internalist coherentism here is that if coherentism and strong internalism were both correct, I think it would be safe to say that no judgments about phenomenology (or anything empirical, for that matter) could be justified by a being with anything like human cognitive capacities, regardless of what the true mind-body theory was. (The requisite abilities to recognize inferential relations between judgments—and standing beliefs—would simply be beyond the grasp of human intellectual powers.) Hence, epiphenomenalism would not be at a disadvantage.

> (Strong Internalist Foundationalism) A judgment about phenomenology *p* is justified iff (1) the agent making *p* is possessed with a good overall reason *r* for thinking *p* is probably true, (2) possessed with a good overall reason *s* for thinking that *r* is a good overall reason for thinking *p* is probably true, (3) the agent bases *p* on *r*, and (4) the agent bases the judgment *r* on *s*.

By way of clarification, let me offer a few comments on this definition. First, it should be noted that even though this is a strong internalism (it puts rigorous requirements on the agent to have reasons for her judgment, and to base the judgment appropriately on the reasons), there are stronger internalisms available. Strong Internalist Foundationalism requires only that the agent be in possession of a reason for the judgment, and a reason for accepting this reason. This involves a single level of "justificatory ascent"—the only hypothetical challenges that must be defended against explicitly with reasons are challenges to the judgment itself, and challenges to the agent's reply to those initial challenges. (Any reason possessed by the agent that survives this second round of hypothetical skeptical challenge is *ipso facto* foundational.) But clearly, more rigorous requirements could be formulated—requirements that challenges at this second-level be defended against, or a third, or a fourth. In fact, it could even be required that the challenges be answered until no further challenge can be advanced—until the indisputable foundation of justification is reached. Consequently, it should be kept in mind that we are not considering the most extreme possible internalist theory, but the view we are considering should be extreme enough for our purposes. Second, "reason" need not be understood fully propositionally here. In keeping with the spirit of views like

the one Laurence BonJour has advanced recently,[135] the reasons in question could be pure qualitative properties themselves, or the event that is oneself instantiating them (or something similar in spirit)—these things are not plausibly thought of as themselves propositional in character.[136]

Finally, let me offer a word on what a good overall reason is. I conceive of an overall reason as something relative to a level—something that encompasses all of one's individual reasons at a particular level. What do I mean by 'level'? This is a tricky issue, but I take it that the level of a reason is, in some sense, its "distance" (presumably inferential, or something closely related)[137] from one of the foundational reasons. Thus, having a good overall reason at a level amounts to being such that all of one's reasons at that level sufficiently probabilize the judgment whose justification is in question. The point of the overall reason requirement is to ensure that the dialectic with the imaginary skeptic is not trivially reduced to a single requirement—having sufficiently good reason for the primary judgment. Such an account, while perhaps accurately capturing the internalist position in a sense, would disguise most of the subtle requirements that the internalist wishes to advance.

Now, to fully appreciate the prospects of justification for phenomenal judgments on this internalist view (for any mind-body theory), we must have a good idea what these judgments are like (i.e., what is their content), and what constitutes them (i.e., what phenomenological components express the concepts that comprise the judgment). Both of these questions are (surprise, surprise!) enormously difficult, and have perplexed

---

[135] As in BonJour (1999).
[136] Though no doubt there is much more to be said here.
[137] I avoid saying 'inferential' *simpliciter* here because there is some sense in which the relevant sort of phenomenal judgment is not itself inferential, and so is thus not inferred from other judgments or reasons (more broadly construed). But it seems as though it could still be based on these other reasons in some sense. In any case, this is an issue to be saved for more systematic treatment elsewhere.

philosophers continually since at least the early modern period (and careful psychologists as well since the beginnings of psychology as a formal science).

Thus, for the time being I won't try to explore all the various options, but I will try to make a few remarks appropriate for the preliminary investigation being carried out. Recently, a number of accounts of introspection and introspective judgment have been advanced that place emphasis on the phenomenal state that constitutes the introspecting or the introspective judgment "embedding" the phenomenology of the original state the introspection is directed on. Alternatively, the phenomenology of the original is sometimes described as "constituting" the concepts employed in the introspective judgment.[138] If there is some sort of embedding or similar relation between the judgment and the phenomenology the judgment is about, then there will presumably be no causal relationship between the original phenomenology and the judgment about it. This would be the case even though the judgment would presumably be based on the original phenomenology, and the original phenomenology would constitute the good overall reason for the judgment.[139] (There is, of course, the issue of the requirements of "justificatory ascent" and whether they would be met by this sort of basing relation, but there seems to be no obvious reason why the fact that the basing relation was non-causal

---

[138] Most prominent among these have been accounts given by Brie Gertler and David Chalmers. (The "embedding" formulations are due to Gertler, the "constituting" ones to Chalmers.) See Gertler (2001) and Chalmers (2003). See also Chalmers (1996), especially Chapter 5, for an earlier, less developed defense of a similar view. It should be pointed out that Chalmers delineates between several different varieties of phenomenal concept, not all of which are constituted by the phenomenology they are about (by 'constituted', Chalmers seems to mean something like what I have called 'expressed').

[139] Even if the relevant phenomenal concepts (like PAIN, e.g.) are constituted by/embed the phenomenology they are about, there may be other concepts involved in the judgments that do not. What these other concepts are, and what phenomenology expresses them, will depend on specific details of preferred views on the nature of these judgments (what exactly the propositions involved are), as well as other side issues. I won't try to speculate on the impact these peripheral considerations might have on the justification of the judgments. I'd imagine they probably have little, but I can't defend that speculation here.

would present a novel obstacle where this was concerned.)[140] So if these recent accounts of introspective judgments are correct, then seemingly, if epiphenomenalism were true, the agent would fare just as well in justifying his judgment as if the physicalist alternative were true, since causation would not be part of the crucial basing of the judgment on the reason.

Even if these non-causal accounts are incorrect, there is hope for the epiphenomenalist. (And as mentioned above, Chalmers at least thinks there are classes of phenomenal judgments—even ones that employ what he calls "pure" phenomenal concepts, not in any way dependent on publicly accessible criteria—whose salient concepts are not constituted or expressed by the phenomenology they are about.) If the phenomenology that expresses the crucial phenomenal concepts (like PAIN, e.g.) is something like a remembered instance of pain, or an imagined instance of pain (somehow colored by memories of pains previously had), its neural basis may bear enough structural similarity to the neural basis of the presently had pain (assuming there is such a presently had pain) for the judgment to count as being based on the original phenomenology in the appropriate way (all without there being any qualitative events causing any other qualitative events or anything else). Perhaps even its own tokening would be causally dependent on the neural basis of the pain. (I am assuming epiphenomenalism here, of course, and its concomitant dualism.)

---

[140] Philosophers like Russell and Richard Fumerton, who have defended the claim that there is a special sort of relation between subjects and their experiences that potentially ground various justificatory relations (which they call 'acquaintance'), often claim in addition that we are "acquainted with acquaintance," partly in order to satisfy something along the lines of these ascent requirements. See, for example, Russell (1910) and Fumerton (1995). I have avoided the terminology of 'acquaintance' here because it (and the concept associated with it, for that matter) is quite vexed, and the word has often been used in confusing (and sometimes confused) ways.

But there is one more consideration that I think is the lynch pin in the argument for equating the justificatory prospects of phenomenal judgment on epiphenomenalism and physicalism where strong internalism is concerned. I have taken it as axiomatic throughout this entire dissertation that one cannot obviously, pre-theoretically verify that epiphenomenalism is false simply by introspection. If all the non-epiphenomenalist mind-body theories (if true) provided some magical experiential cue that indicated epiphenomenalism was not the case, then there would be no need for arguments against epiphenomenalism. One could simply introspect and make a phenomenal judgment, locate the cue, and be done with epiphenomenalism. (Assuming that one's introspective judgment was appropriately justified and so forth.) I take it that no one has managed to convincingly demonstrate the existence of such a cue, hence the indirect ways people have typically attempted to dismiss epiphenomenalism (e.g., as weird, repugnant, unparsimonious, etc.). But if there is no way of "telling from the inside," so to speak, whether epiphenomenalism is the case, and internalism requires judgments to be based on good overall reasons in order to be justified, then a dilemma arises for someone attempting to elude the conclusion that a judgment would fare differently depending on whether epiphenomenalism or physicalism was the true mind-body theory. Either there is something about epiphenomenalism that, if true, stands in the way of a phenomenal judgment's justification or there is not. If there is nothing about epiphenomenalism that, if epiphenomenalism were true, would be an impediment to justification, then the possibility of epiphenomenalism is unproblematic for justification. If, on the other hand, there is something about epiphenomenalism that, if true, would stand in the way of justification, then that is something that stands in the way of justification *on any mind-*

body theory, even physicalism. This is because, since epiphenomenalism can't be automatically ruled out by introspection and phenomenal judgment based on introspection alone, the mere possibility of its being true and the agent's inability to rule it out using the tools internalist foundationalism provides is enough to stand in the way of justification no matter what.[141] So, either way, the question of whether a judgment is justified according to strong internalism is not dependent on what the true mind-body theory is. In a way, this isn't terribly surprising. Since strong internalism makes justification a matter of things directly accessible to the agent, and the truth of epiphenomenalism is not something directly accessible to the agent, it is no shock that the truth or falsity of epiphenomenalism does not impact the answer to the justification question (at least not directly—it could of course indirectly impact it by, as just discussed, making all phenomenal judgments unjustified, owing to the open epistemic possibility or decent epistemic probability, from the agent's perspective, that it is in fact the case).[142]

Let us now move on to consider the second of the four justificatory theories—a moderate one inspired by the account Chalmers himself gives in Chalmers (2003). I formulate it as follows:

---

[141] If the agent could marshal reasons for supposing epiphenomenalism very unlikely, then this might change matters, but it is very difficult to see how the agent could be in possession of such a reason given the meager tools that are available.

[142] One could, of course, think that the falsity of physicalism is knowable *a priori* and so directly accessible to the agent in the relevant sense, and thus rule out at least one of the options in the debate before even considering the evolutionary argument. This is an important point, and one well worth reflecting on. I will discuss it once our treatment of the evolutionary argument has been expanded to include examination of interactionist hypotheses.

(Moderate View) When a subject forms a phenomenal judgment *p* introspectively, then *p* is *prima facie* justified.[143]

What is it for a judgment to be '*prima facie* justified'? It is perhaps easier to answer this question indirectly, by tackling a different one—namely, what is required for a judgment's *prima facie* justification to amount to justification *simpliciter?* I take it that what turns *prima facie* justification into justification is the absence of any defeaters. (Someone sympathetic to a defeater sort of view could then claim that all and only *prima facie* justified phenomenal judgments, where no defeaters are present, are justified phenomenal judgments.) What, then, is a "defeater"? There are numerous ways to cash out the notion of a defeater, some more internalist and others more externalist (thus providing a continuum of possible moderate views). A more internalist variant might require that, if a skeptical challenge occurs to the agent (whether spontaneously, pointed out by another agent, etc.), she provide a defense against it.[144] (The arising of the skeptical challenges, and failure to meet them, would constitute the individual defeaters.)[145] A more externalist variant might require instead that certain impediments

[143] I do not use Chalmers's own formulation for several reasons. First, his terminology differs from ours—he, for instance, often uses the word 'judgment' to mean something very different from what we mean by it, at least in much of his work. (He uses the term to pick out something subconscious.) Second, he doesn't intend the thesis to apply to all phenomenal judgments arrived at introspectively using pure phenomenal concepts. He proposes that at least some judgments (i.e., what we call 'judgments') involving standing phenomenal concepts are justified in other ways. Third, he employs the notion of ACQUAINTANCE, which as I noted above is an extremely vexed and difficult notion that is best avoided for our purposes.

[144] Some might prefer to call the challenges themselves (whether answered or unanswered) 'defeaters', and then rework the formulation of the condition that moves a judgment from *prima facie* justification to justification, having it claim that there be no "undefeated defeaters" instead of no defeaters at all. This is really just a terminological issue, so the reader can feel free to substitute such an account if it is more comfortable.

[145] Various caveats might be provided to this initial formulation to ensure (e.g.) that previous negligence has not created in the agent the habit of ignoring or failing to recognize pertinent skeptical challenges. (Notice, in fact, that this qualification could be made so strong that the moderate view would collapse into something close to the strong internalism discussed earlier.) It is interesting to recognize that Gettier cases

to reliable judgment be absent, whether or not those impediments were cognitively accessible to the agent or things that had occurred to the agent to rule out. (These might include inattentiveness, drug use, extreme specificity in the content of the judgment, etc. The presence of any one of these things would constitute a defeater.)

If, to examine a possibility already introduced in connection with strong internalism, the phenomenology the judgment is about is somehow embedded in or constitutes the phenomenology that expresses the judgment, then once again epiphenomenalism will fare no worse than the physicalist alternative.[146] This is because presumably concerns of mental causation will have no bearing on whether a defeater is present if judgments embed the phenomenology they are about. (If this last claim is false, then it is the job of the opponent of the epiphenomenalist to explain why. For our preliminary purposes, we needn't engage the anti-epiphenomenalist further on this matter.) If the embedding style view is mistaken, though, and the phenomenology that constitutes the judgments is wholly separate from the original phenomenology, then epiphenomenalism may be at a disadvantage. This is because the original phenomenology will be unable to cause the judgment.[147] (And this failure to cause the

---

might be common with this sort of view, assuming that the negligence avoidance qualifications weren't made too strict. Imagine a situation where failing to attend to one's phenomenology very closely when introspecting did not itself create a defeater in typical cases. Then, judgments formed on the basis of such lackadaisical attendings could on occasion be both justified and true, but fail to count as knowledge, since their truth was an accident. (In nearby possible worlds, the same judgments would often be false, and would not effectively "track" the truth.)

[146] Incidentally, if the reader is wondering how inattentiveness would be possible on this sort of view of introspection and introspective judgment, it is indeed difficult to say, and this may be a serious weakness of theories like Chalmers's. This issue is an extremely difficult one, and consideration of it must be set aside.

[147] Assuming again that qualitative events do not cause other qualitative events if epiphenomenalism is the case. Although this is not entailed by the definition—the definition entails that qualitative events are causally inert only with respect to physical events—I defended the plausibility of this assumption earlier on.

judgment might itself constitute a defeater, since it might be thought to introduce the possibility of serious unreliability in phenomenal judgment.)[148]

But is it really plausible to suppose that a lack of both embedding and causation between original phenomenology and judgment, even if it were present, would be problematic for epiphenomenalism? Not clearly. In fact, I would say far from clearly. Presumably there would be some causal relationship between the neural bases of the original phenomenology and the judgment,[149] and this would presumably be enough to ensure reliability just as well as direct causation from phenomenal to phenomenal. (And the agent could be aware of this indirect causation's tendency to preserve reliability, and hence use this awareness to head off defeaters on the more internalist versions of the moderate view.) Moreover, if the neural events that wound up constituting the neural basis of the judgment and were caused by the neural basis of the original phenomenology tended to be structurally similar to the neural basis of the original, that would be a further plus for epiphenomenalism. This would provide good grounds for supposing that the phenomenology of the judgment was mirroring (in some sense) the phenomenology of the original judgment. If the phenomenology that expressed the salient concepts in the judgment had affinities with the phenomenology that the concepts purported to represent, a proposal many have thought plausible for phenomenal concepts (including, of course,

[148] On the more internalist view, it would be the skeptical possibility of the original phenomenology's failing to cause the judgment appropriately that would threaten to serve as the defeater if it couldn't be effectively answered.

[149] This might be one of the key factors making it the case that the judgment is in fact about the original phenomenology.

Chalmers), then the mirroring would be a good indication that the original phenomenology was satisfying the concepts, and making the judgment true.[150]

In any case, even if my opponent is not satisfied either that there is an embedding/constitution relation between original phenomenology and judgment phenomenology which makes epiphenomenalism irrelevant to justification, or (in case there is no such relationship) that the failure of causation between phenomenal events does not impede the prospects for justification, it is her job to discharge the burden of proof. She must explain why the lack of causal connection is so problematic for epiphenomenalism on one of these moderate theories, since I have made it clear that, on a preliminary inspection, there is no obvious reason to suspect epiphenomenalism of creating special justificatory difficulties if true.

In addition, on the more internalist of these moderate options, we run into an analogous issue as with strong internalism. Even if my arguments above all have false conclusions, and epiphenomenalism really would present an obstacle to justification if it were true, then just the fact that the agent is not directly aware of the falsity of it is enough to create an obstacle for justification, no matter what the true mind-body theory is. This is because presumably the looming possibility of epiphenomenalism, a possibility that cannot be ruled out (or verified to be improbable) by the agent, will constitute a defeater.[151] Thus the truth of epiphenomenalism would not result in the

---

[150] Obviously it would matter a great deal what the structural similarities were exactly, but tackling that sort of issue is far beyond the scope of this work, and indeed, probably beyond the capacity of present cognitive science to address in much empirical detail.

[151] If the agent is not sufficiently reflective to recognize this skeptical challenge (and has not been negligent heretofore in a way that would make his failure to be sufficiently reflective problematic), then some versions of this moderate internalist view would countenance the justification of the judgment (as discussed in a footnote above), but only at the cost of permitting Gettier cases to abound. (They would count the judgment as justified because no skeptical challenge came to the attention of the agent.)

prospects for justification being any different from what it would be on physicalism (on a view of epistemic justification like this).[152]

Now that I have completed my preliminary discussion of both the strong internalist and the moderate theories and found neither of them especially hostile to the prospects of justification of phenomenal judgment on epiphenomenalism, let me transition into a discussion of the more externalist options—process reliabilism and a counterfactual dependence theory.

Neither sort of view can be generally said to be more externalist than the other—they are basically just alternative externalist paradigms for conceiving of epistemic justification. Thus, I will begin with process reliabilism, and transition into the counterfactual dependence theory, but nothing about this order should be understood as implying anything about the views' respective degree of externalism.

Process reliabilism has a long and storied history in analytic epistemology as a theory of justification of empirical beliefs (about physical objects, etc.), but to my knowledge it has not often been applied to phenomenal judgment.[153] But there is no reason in principle why it cannot be so applied, and I will do so here.

---

[152] The reason a similar problem doesn't arise on the more externalist of these moderate theories is that these theories, on account of their externalism, can stipulate that the truth of epiphenomenalism would constitute a defeater (and do so in a principled fashion), but that there would be no analogous defeater if physicalism were true. (The externalism helps because it obviates the need for the defeaters to be based on things the agent has direct access to, and the truth of epiphenomenalism is not something the agent has direct access to.) Notice too that there are further subtle distinctions in moderate views that could be drawn here as well—a view could specify different standards for what counts as a potential defeater and what counts as a consideration canceling the challenge. For example, someone might tend in an externalist direction where potential defeaters are concerned (i.e., that they need not be noticed by the agent), but tend in an internalist direction where canceling their threat to justification is concerned (i.e., the challenge would have to be overcome by the agent's use of reasons cognitively accessible to her).

[153] Perhaps the classic statement of process reliabilism is in Goldman (1976). The view has gone through countless incarnations and manifested itself in numerous prominent variations, however, so no individual version can claim to be paradigmatic. Incidentally, Chalmers argues in Chalmers (1996) that something along the lines of reliabilism could not be the correct theory of justification for phenomenal judgments. His arguments were criticized by Tim Bayne in Bayne (2001), and I believe Chalmers no longer stands by

There are many complicated formulations of process reliabilism designed to circumvent various preliminary objections and counterexamples to the view. For our purposes, I will simply articulate a basic version, with the understanding that appropriate bells and whistles can be added as needed to deal with familiar difficulties (none of which are germane to the main issue at hand). Here is how we will formulate it:

> (Process Reliabilism) If an agent's phenomenal judgment *p* results from a reliable process, and there is no reliable process such that, had it been used in addition to the process actually used, would have resulted in the agent's not making *p*, then *p* is justified.

As before, let us consider in turn the possibility that the phenomenology of the judgment is partially constituted by/embeds the original phenomenology, and the possibility that it does not. If the phenomenology that expresses the relevant concepts in the judgment is constituted by the original phenomenology, then it is very plausible that judgments of the sort will be reliable (perhaps even 100% reliable, since after all it seems there is no way for the judgment to be expressed and also false). In addition, it would be very implausible to suggest that some other reliable process type would have changed the judgment and made it false if it were in fact used. Thus, the second part of the above theory of reliabilist justification is also satisfied. And this reliability will be consistent

---

those original arguments. (He now prefers to make use of a consitution relation between original phenomenology and judgment phenomenology, as discussed earlier, to account for justification. Before he appealed strictly to acquaintance with experience to do all the difficult work, only very suggestively floating the possibility of a constitution relation. (See Chalmers (1996), pp. 203-208 for discussion.)

regardless of whether epiphenomenalism or physicalism is true, since the reliability will not be grounded in any causal relationship between qualitative events.

If, on the other hand, the phenomenology that expresses the relevant concepts in the judgment is not constituted by the phenomenology of the original, does epiphenomenalism present an obstacle for justification? Again, for similar reasons as we saw above in the discussion of the moderate options, there is no reason to suppose so. If there is a predictable causal pattern between the neural basis of the original phenomenology and the neural basis of the judgment, then why suppose that the reliability of the judgment is more suspect than if the causal relationship were directly from phenomenology to phenomenology (as would be the case if physicalism were true, presuming it was plausible to understand the physical to physical causation involved as in some sense embodying the phenomenological causation[154])? Why would a direct causal relationship be more likely to produce reliability than an indirect one (via physical neural bases) grounded in stable natural laws? I don't see why. Consequently, at this stage I don't accept that epiphenomenalism would provide an obstacle to justification if process reliabilism is in fact true.

Our final representative theory to examine is a counterfactual dependence theory. This theory, much like process reliabilism, has received numerous formulations, and if anything typically contains even more bells and whistles to deal with counterexamples and *prima facie* difficulties than its reliabilist cousin.[155] (Once again, as with reliabilism, I will offer a basic formulation, and leave the technical details—motivated largely by

[154] As I alluded to above, showing this is not a trivial metaphysical undertaking. But taking on general issues of mental causation and property individuation that stem not from robust dualism but from concerns about multiple realizability and determinate/determinable relations is far outside the scope of the present work. For such a discussion, see Yablo (1992).

[155] A far as I know, the theory was first formulated in Nozick (1981), and the discussion there remains a classic one.

unrelated problems—for elsewhere.) The following will be our formulation (inspired by Nozick's own):

> (Counterfactual Dependence Theory) A judgment that *p* is justified iff (1) if *p* were the case, the agent would judge that *p* (if the agent judged one way or another on the matter), and (2) if *p* were not the case, the agent would not judge that *p*.[156]

Let us then consider how phenomenal judgments would fair on epiphenomenalism in comparison with physicalism. If there is an embedding or consitution relation between the phenomenology of the original and the phenomenology that expresses the judgment, (1) and (2) will hold on either epiphenomenalism or physicalism. The reason for (1) being satisfied is that, since the original phenomenology constitutes (some of) the relevant phenomenology that expresses the judgment, if the original phenomenology were different, then the judgment would be different as well (since the phenomenology to express the original judgment would not be available). And the reason for (2) being satisfied is similar. The agent's judgment will be constituted by/embed the phenomenology of the original, so there will be no way for the agent to judge falsely (since a difference in the original phenomenology will prevent the judgment from being made in the first place). Because *ex hypothesi* phenomenal to phenomenal causation is not involved in this process at all (and certainly not phenomenal to physical

[156] The parenthetical remark is included, since it doesn't seem required for justification that if *p* were the case, then the agent would judge that *p*. What seems to be of primary importance is that the agent would not judge falsely that not *p* if *p* were the case.

causation either), there is no opportunity for epiphenomenalism's distinctive aspects to get it into trouble.

If, on the other hand (to traverse similar territory), there is no constitution/embedding relation, for similar reasons as with process reliabilism there is no special grounds for the epiphenomenalist to worry. Epiphenomenalism will presumably posit lawful causal relationships between the neural basis of the original phenomenology and the neural basis of the phenomenology that expresses the judgment. These lawful causal relationships will be such that they can support counterfactuals—if the neural basis of the original doesn't hold, neither will the neural basis of the phenomenology of the judgment; and if the neural basis of the phenomenology of the judgment holds, presumably it will have been caused by the neural basis of the original phenomenology. In any case, even if these causal relationships fail to support counterfactuals in the right sort of way, it is not at all clear that phenomenal to phenomenal causation on physicalism would do a better job of supporting the salient counterfactuals. (This is analogous to a point made with respect to process reliabilism.) Thus, epiphenomenalist phenomenal judgments don't seem to fare any differently if the counterfactual dependence theory is true than they do if any of the other theories already considered are.

Clearly, much remains to be said about these thorny epistemological issues, and no brief treatment can do them full justice. But I think I have said enough in this space to convincingly head off preliminary concerns about the justification of phenomenal judgments if epiphenomenalism is true, and consequently our ability to use them as evidence. (Or more carefully, said enough to head off preliminary concerns that the

prospects of phenomenal judgments' justification is adversely affected by the truth of epiphenomenalism.) This is all I have set out to do in my limited inquiry here.

To recap briefly, in this section we considered four different paradigmatic theories of justification—strong internalist foundationalism, a moderate defeater theory, process reliabilism, and a counterfactual dependence view. At least preliminarily, I have found that epiphenomenalism shows no signs of causing particular difficulty for epistemic justification on any of these theories. In each case, this is for at least one of two general reasons: either the truth of epiphenomenalism poses no problem for justification according to the view in question, or, if the truth of epiphenomenalism does pose problems for justification, similar problems arise even if physicalism is true.

VI. The Connection Between Phenomenology and Behavior

We come now to the final objection to the anti-epiphenomenalist evolutionary argument—objection (H). Recall that this objection challenges the argument's use of the assumption that there is an appropriate "unbreakable" connection between phenomenology and behavior if physicalism is true.

Let me explain in a little more detail what this worry amounts to (or at least one version of the worry—we will see coming up that the worry comes in a number of different forms). Throughout the argument, it was assumed (with minor qualifications) that there are few constraints on what the possible bridge laws could be from neural bases to phenomenology according to epiphenomenalism. (Indeed, this assumption was defended in a response to objection (C)—the objection that claimed that we would have background reason to expect the phenomenology associated with avoidance behaviors to

be negative on any mind-body theory.) This is what allowed the argument to conclude that the correlations between INP and distal stimuli were surprising on epiphenomenalism in the first place, since there were a wide variety of possible (and not especially improbable) epiphenomenalist bridge laws that would not have maintained the correlation, but still would have preserved the organism's appropriate behavior in response to the dangerous stimuli. Recall, though, that an important component in the overall strategy of using this information against the epiphenomenalist was that physicalism leads us to expect something different—in particular, that physicalism leads us to expect the very correlations we did find. *But* after all, someone might contest, physical to physical causal laws are themselves contingent (on typical construals), so why suppose that they would be exempt from the same kinds of worries? But if they are not exempt, then the argument falls apart, since neither of the competing alternatives leads us to expect the correlations we actually find, and so neither is confirmed or disconfirmed (appreciably, at least) by the evidence.

This is indeed a deep worry, and one that cannot be adequately examined in a brief discussion. Ultimately, as I mentioned above, I think the insight embodied in objection (H) ultimately dooms the original argument, but unlocks the door to appreciation of the true role of empirical and conceptual considerations in productive debate about the mind-body problem. In fact, it also provides us with a natural segue into a discussion of what has up till now been the lone spectator among the main mind-body theory competitors—interactionism. (It provides the natural segue because opening up issues about the connection between phenomenology and behavior provides a good opportunity to discuss the overall outlook of all the major theories, since all have

distinctive views on the subject. Heretofore, it has made the most sense to restrict our attention to just epiphenomenalism and physicalism, since the objections considered applied primarily to epiphenomenalism—or were best discussed in the epiphenomenalism context—with physicalism a useful foil.)

As a result of the systematicity of the considerations we will need to explore, a discussion of this objection to the argument must wait for a future chapter. This chapter is coming up shortly, after a brief interluding chapter that discusses alternative formulations of evolutionary-style arguments. (Some of the material in the upcoming chapter will also point us naturally in the direction of our culminating discussion.)

For the moment, though, let me briefly recap the territory we have traversed in this lengthy chapter. In it, I presented and replied to seven overall objections to the anti-epiphenomenalist argument formulated at the beginning of the chapter. (And presented, but did not reply to, the eighth and final objection.) I found all of these objections wanting. Some, such as (A) and (C), I found to be obviously mistaken, while others, such as (D) and (G), I found interesting and suggestive, but nonetheless implausible at this stage. They would need to be developed in much more detail to have any hope of being persuasive.

Chapter 3—Alternative Formulations of the Evolutionary Argument

At several points in discussions above, I alluded to the possibility of providing alternate formulations of the evolutionary argument against epiphenomenalism (or, to put the point more accurately perhaps, formulations of other arguments similar to the evolutionary one we have already seen). My task in this chapter will be to make good on the suggestive remarks above and actually outline ways that the argument could be changed or reformulated, and also to discuss the dialectical relevance those changes would have.

There are three variations worth discussing. These are (in the order in which they will be examined):

(1) Changing the correlations used as evidence from phenomenology/distal stimuli to phenomenology/behavior (or adding these further correlations in as evidence, if you like).

(2) Dispensing with evolutionary evidence and replacing it with evidence about the long-term survival of presently living persons (or those of recent history).[157]

(3) Using more precise phenomenological evidence. (E.g., rather than just the evidence that a particular stimulus is associated with an INP, the evidence that it is associated with a very specific burning phenomenology.)

Ultimately, I will conclude that these alterations in the basic argument are listed in increasing order of dialectical significance. (1) has either no or virtually no impact on the

[157] This is the variation (in a first-person form) alluded to above in replying to the objection which contended that we cannot legitimately use the correlations between other people's phenomenology and the distal stimuli that they are exposed to as evidence.

force of the argument one way or another, especially if we admit reasonably fine-grained phenomenological evidence (in the spirit of (3)). (2) has very little—it may weaken the strength of the argument's conclusion slightly (owing to the reduction in the breadth and scope of evidence the argument can appeal to), but this weakening is barely noticeable. (It has the corresponding advantage, however, of using less controversial evidence, albeit only ever so slightly less controversial.) (3), on the other hand, has the potential to significantly strengthen the prospects of the argument. Indeed, it is consideration of (3), in addition to reflecting on objection (H) from the previous chapter (the objection which questioned the "unbreakableness" of the phenomenology/behavior connection on physicalism), that will ultimately propel us toward consideration of the overall prospects for the argument, pursued in the next chapter. (For the time being, however, we will not be pursuing or answering objection (H). Those issues will be taken up again in the next chapter. For the purposes of this chapter, I will take the primary outstanding assumption of the anti-epiphenomenalist evolutionary argument—the one criticized in objection (H)—for granted. The discussion here will evaluate the relevance of the changes given that we do accept this central assumption. If we do not, then there will be no pro-physicalist conclusion to strengthen or weaken with the currently considered variations, since the argument will fail to lend any support to physicalism, regardless of the niceties of formulation.)

I. Phenomenology/Behavior Correlations

Let me begin, then, with a consideration of (1)—changing the evidence used from information about the correlations between phenomenology and distal stimuli (e.g., INP

with cuts to the skin) to information about correlations between phenomenology and behavior (e.g., INP with avoidance behaviors). One initial thing to notice about an argument that employed phenomenology/behavior correlations instead of phenomenology/distal stimuli correlations is that it would seem *prima facie* that the argument would be harmed as a result, since there would be far less evidence to go on. Rather than having a whole list of correlations (INP with burning, INP with bee stings, INP with cuts to the skin, etc.), the argument might be reduced to a very small list, since INPs produce only avoidance behaviors in the general case. (I.e., there is no specific kind of avoidance behavior that INPs produce as a general rule—a burning sensation in the hand produces one kind of avoidance behavior, a bee sting in the leg another.)

There are numerous ways to sidestep the worry that the difference produced is anything but superficial, though. The most straightforward (and to my mind the best) way is to claim that the method of formulating evidence that results in an apparent mismatch between the evidence in the two situations is based on an unprincipled difference in the level of fine-grainedness of evidence admitted in the respective scenarios. In the phenomenology/distal stimuli case, although the phenomenological evidence was not terribly specific (INP vs. IPP vs. neutral phenomenology, perhaps), the distal stimulus evidence was. The proponent of the argument was not required to restrict herself to noting only whether a stimulus was survival-threatening or survival-conducive (or neutral)—she could employ information about the specific nature of the stimulus. But in the phenomenology/behavior scenario, although the standard for fine-grainedness of phenomenological evidence is the same, the standard for behavioral evidence is much more general than for distal stimulus in the previous example. We are only allowed to

use the information that the behavior was an avoidance or a seeking out (or neither), not anything about the specific nature of the behavior. Once we are allowed to register more fine-grained behavioral information, the equilibrium is roughly restored.[158] (I say "roughly" here because there may still be some minor discrepancies. For example, take the distal stimulus correlations for a moment. If I am stung in the arm by a bee, I will have a certain specific phenomenology, and ditto if I am cut in the same place by a knife. So there will be two clearly different pieces of evidence here intuitively. But there may not be two clearly different pieces of evidence in terms of behavior correlations, because there may not really be two distinctively different behaviors that result in the two cases, and so the evidence in both may simply be INP with withdrawal of the arm. Once we explore the use of finer-grained phenomenological evidence, in connection with variation (3), even these minor discrepancies disappear, and the evidence is once again completely isomorphic.)

Now that we have settled the preliminary matter, it is on to the main issue of arguing that one can employ either slate of evidence (or both) without affecting the force of the argument appreciably one way or another. The key point to notice (one that has come up before) is that natural selection selects for behavior; if an organism does not respond to a given stimulus with the correct behavior, then its prospects for survival will diminish, perhaps substantially. And given that all the organisms around now—in particular, human organisms—are members of species that have evolved (typically in large part due to natural selection), so long as they are phylogenically normal they will be

[158] In both the behavior and distal stimulus scenarios, the correlations are only one directional when the phenomenological evidence is kept general but the other component allowed to be formulated in specific detail—e.g., "if knife cut to the arm, then INP," or "if specific kind of arm withdrawal, then INP." Issues about the generality of the phenomenological evidence will be taken up in connection with variation (3).

guaranteed to have very predictable correlations between exposure to most dangerous stimuli and avoidance behavior (and the same *mutatis mutandis* for helpful stimuli).[159] And this will be true regardless of the true mind-body theory (since after all that true mind-body theory brought the species from its sentient origins to where it is today). Consequently, since (restricting ourselves to the negative case) dangerous stimuli go hand in hand with avoidance behavior, whether the evidence admitted couples phenomenology with stimulus or with behavior, the results will be the same: physicalism will lead us to expect the correlations, while epiphenomenalism won't, for the same familiar reasons—the ones adduced in the original formulation of the evolutionary argument.

So, I conclude (slightly provisional on the discussion of variation (3), for the minor reason about non-isomorphic evidence discussed above) that behavioral evidence can be substituted for evidence about distal stimuli with essentially no effect on the argument. Thus, I will continue to use distal stimulus style evidence (since that is what we have been working with thus far), but it should be understood that an argument that employs the relevant behavioral evidence instead is no less plausible.

II. The Survival of Presently Living Persons

Let us continue on then to a discussion of (2)—the variation that uses evidence about the long-term survival of presently living persons (or human beings in recent

[159] Although, of course, many helpful stimuli are not as dramatically helpful as dangerous stimuli are dramatically dangerous. This could open the door for some less adaptive responses to helpful stimuli, and also introduce another subtle way in which (1) could have dialectical significance—because sometimes helpful stimuli generate INPs (because our cognitive systems are risk averse), the correlations between stimuli and phenomenology might be less "pure" than correlations between behavior and phenomenology. But I take it that this difference would be tiny enough that it is not worth exploring further here. For an interesting discussion of a variety of tangentially related issues, though, see Stich (1983).

history) rather than evolutionary evidence (i.e., evidence about the survival and development of the human species over many generations due to natural selection).

Given that evolutionary evidence of the general sort being used here is fairly uncontroversial (one might even call it background knowledge, rather than evidence), why might someone want to dispense with it and use thinner evidence? The most sensible reason would be if the individual had a theory of evolution that placed little emphasis on natural selection (as opposed to other kinds of evolutionary pressure). (Though even this reason is still a bit farfetched, I admit.) In this situation, the adaptive advantage appropriate dispositions toward harmful stimuli would play would be minimal, and so the likelihood that evolution itself would produce them wouldn't be especially high. And if this were the case, the evolutionary price to be paid by organisms (and species more widely) for having inappropriate dispositions in response to stimuli wouldn't be great, and so wouldn't itself contribute much to ruling out mind-body theories that resulted in such inappropriate dispositions.

The good news is that even someone in this position can employ what is, by his own lights, appropriately conservative and palatable evidence, and still get a conclusion that is nearly as strong as the one from the original argument (or so I will contend). As I alluded to a moment ago, this evidence is the evidence employed in (2).

The idea is that we do not need to know anything about evolution to establish that there are important correlations between phenomenology and distal stimuli that are expected on physicalism (again, granting it the controversial assumptions about phenomenology and behavior that will be investigated later) but not on epiphenomenalism. Recall that the evolutionary evidence establishes that if organisms

were disposed to seek out too many things that placed their survival in jeopardy, or were disposed to avoid too many things that enhanced their prospects for survival, they would be likely to die and fail to reproduce, with the obvious consequence that their genes would not be passed on to future generations. Thus, according to the argument, the implication would be that organisms around today, since they would inherit the traits of their successful ancestors, would be likely to share their ancestors' appropriate dispositional capacities.[160] But then the impressive match between their phenomenology (INP and IPP) and the distal stimuli would be a surprise if epiphenomenalism were true, since phenomenology would have no causal impact on behavior or play any role in grounding behavioral dispositions.

But do we really need any of this fancy evolutionary evidence? If the basic insight of the argument is correct, wouldn't it be enough if all I knew was that I, an adult human being, was still around after decades of negotiating environmental challenges to my survival? What person has not come across situations at some points in his life where he probably would have perished if not for the presence of a fortuitous behavioral disposition? If I were disposed to drink large quantities of bleach as a child (perhaps because it tasted sweet) or place my limbs on hot stoves (maybe because the stoves felt soothing to the touch) or pour boiling water on myself, more than likely I would have died by now. The same goes for most other people, I would imagine. (The sad fact that subjects who have congenital pain insensitivity tend to live very short lives is testament to this. Although congenitally pain insensitive individuals do not have mixed up

[160]Though I will not address them here, some philosophers have presented arguments that the close association between pain/pleasure and reproductive/survival success is in fact evidence for natural selection, since alternative theories of the development of conscious life would lead us to expect differences from this pattern. See, for instance, Draper (1989) and (1997).

phenomenology—they have affectively neutral experiences in place of normal ones, like feeling pressure instead of pain—because of physiological abnormalities in their nervous systems they are not disposed to avoid harmful stimuli.)

The bottom line is that I recognize that I *am* disposed to avoid most things that are harmful to my prospects of survival and disposed to seek out most things that are helpful to them, and I recognize that the same is true for most of the other people around me. And I also recognize that there is a smooth correlation between certain kinds of phenomenology and those behavioral dispositions.[161] If epiphenomenalism were the case, there would be no reason to expect those very smooth correlations. Although evolution does a good job of accounting for why I (and others) would luckily wind up with the dispositions I (and others) have (i.e., because if my ancestors or their ancestors didn't have them, those ancestors would have died out and we would not have existed), nothing about the evolutionary evidence is especially important to evaluating the smoothness of the correlations.

Granted, taking into account the evolutionary evidence doesn't hurt (so long as we are prepared to admit that the salient evolutionary claims really are true). Maybe some people aren't confident in their grasp of their own dispositions or their ability to discern the dispositions of people around them. Maybe they don't believe they are in fact disposed mainly to avoid harmful things and seek out beneficial things, and so consequently unimpressed by the smooth correlations they find between phenomenology and distal stimuli. (In the first-person instance, they might even figure phenomenology is misleading them in lots of cases into having inappropriate dispositions.) Evolutionary

[161] Of course, this is to take for granted that we can answer the earlier-examined epistemological objections to the claim that we know the relevant correlations hold.

evidence could reassure these people that it is likely that their dispositions really are in line with what is beneficial. But it really isn't contributing very much to the argument.

Thus, I conclude that (2) is a fine way of conceiving of and formulating an argument based on the central assumption of the evolutionary one (namely, that there is a direct and special connection between phenomenology and behavior on physicalism, but not epiphenomenalism). For the purposes of this work, I will continue to use the evolutionary argument as the paradigm version, but it should be kept in mind that this alternative is available and seems to work no less well (or just barely less well). It is interesting to note, however, that there seems to be no discussion in the literature of this alternative, nor even acknowledgment that it exists.[162]

III. Fine-Grained Evidence

Now that we have seen the prospects for (1) and (2), let's consider (3)—a version that I claimed above could have much more dialectical significance than the two alternative versions already discussed. (Indeed, it would have much more dialectical significance if the central unexamined assumption of the argument—the one criticized in objection (H)—were correct, and in addition the use of the kind of evidence (3) employs is licit in the circumstances. It would greatly strengthen the pro-physicalist force of the argument.) Recall that this version uses much more precise, fine-grained phenomenological evidence than merely that the phenomenology in question is intrinsically positive or negative (or neutral). Though somewhat more complicated, this way of formulating evidential claims seems natural and principled in the context, since

[162] The same cannot be said of (1), however. Some of the classic formulations of the evolutionary argument, such as James's in James (1890), employ elements of the kind of evidence involved in (1).

(as was discussed above) the heretofore accepted way of formulating the distal stimulus evidence claims is fine-grained. In other words, the accepted way of formulating claims doesn't just register whether the stimulus is survival-threatening or survival-conducive, but rather registers precisely what sort of stimulus it is—e.g., a sharp cut to such and such place on the arm.

One would expect the two sides of the correlation to be similarly detailed and fine-grained, and stepping up the fine-grainedness of the phenomenological side accomplishes this goal. In addition, taking into account fine-grained phenomenological evidence would have the further virtue of employing the most determinate evidence at our disposal (or at least much more determinate evidence than previously), which many philosophers think is a requirement for marshaling evidence as part of an abductive or inductive argument.[163]

What would the effect be on the evolutionary argument of employing this more determinate evidence? Let's examine things on the epiphenomenalist side first. If the phenomenology/stimulus correlations used in the argument were changed from things like "INP with deep knife cuts to the arm (of such and such specific sorts)" to "distinctive INP *P* (i.e., whatever the distinctive negative feel associated with deep cuts to the arm is) with deep knife cuts to the arm (of such and such specific sorts)," epiphenomenalism would seemingly do a far worse job leading us to expect the evidence. The reason is that previously, there were only three possible phenomenologies that could have been associated with the knife wounds. (These were INP, IPP, and neutral.) Thus, even if epiphenomenalism could have produced any of these phenomenologies without changing

[163] White (2000) argues for this principle and provides what he believes are counterexamples to principles that advocate other approaches. I do have sympathy for this use of fine-grained evidence and will employ fine-grained evidence when examining the overall argument in the next chapter.

behavior in the slightest, there weren't that many different ways for epiphenomenalism to produce a "mismatch" between qualitative feel and distal stimulus. If the bridge law between physical/functional and phenomenal had about an equal chance of producing any of these three kinds of phenomenology, then epiphenomenalism's chances of producing any of the actually observed correlations would be around 1/3. (And even if this assumption isn't realistic, then—assuming the earlier replies to objections are on target—the true likelihood wouldn't be terribly different, say by more than a factor of 2 or 3 in either direction.)

What would epiphenomenalism's likelihood be of producing the overall set of actually observed fine-grained correlations? That would depend largely on the extent to which the different qualitative events involved in the respective correlations shared common neural underpinnings. The reason this information is crucial is that, if the neural underpinnings of the various (e.g.) INPs had in response to different survival-threatening stimuli varied greatly, the individual bridge laws from neural base to phenomenology would be likely to be considerably more independent than they would be if the neural underpinnings did not vary much at all. And if these bridge laws were independent, there would be many more opportunities for the bridge laws in different cases to wind up producing mismatched phenomenology—e.g., IPPs with dangerous stimuli—and so a much greater chance that the correlations not line up as smoothly as they in fact do.

But now, it is important to notice that no matter what the answer is to the question about common neural bases, epiphenomenalism will tell us the actually observed fine-grained evidence is far less likely than the actually observed coarse-grained evidence. This is because, for any given distal stimulus, there are presumably countless different

kinds of determinate phenomenology that could be produced by a particular neural basis (as specified by the bridge laws). The neural successor to a knife cut in the arm, for example, would not be restricted to producing merely an INP, IPP, or neutral phenomenology, but could produce any of a large number of detailed phenomenologies, some intrinsically negative, some intrinsically positive, and some neutral. Now (again presuming relative parity between the likelihood of producing the respective phenomenologies), this implies that the likelihood of producing the exact phenomenology actually observed is very low, far lower than 1/3.[164] (And any independence in the correlations stemming from a lack of commonality in the bases of the various phenomenologies will only further exacerbate the problem for the epiphenomenalist. For example, the chances of two independent bridge laws resulting in phenomenology that matches what is actually observed is roughly 1/3 * 1/3 = 1/9 on the coarse-grained style, while on the fine-grained—supposing 1/50 is a moderate or conservative estimate of the number of fine-grained possibilities[165]—is 1/50 * 1/50 = 1/2,500.)

In itself, this outcome of using fine-grained evidence need not result in further disconfirmation for epiphenomenalism vis-à-vis physicalism, of course. This is because, for epiphenomenalism to be further disconfirmed with respect to physicalism as a result of the evidence, physicalism cannot have a proportionally equal or worse loss in its own

[164] It is not terribly far-fetched to suppose that some kinds of phenomenology could not be what is produced by a given neural basis, or at least that some kinds of phenomenology would be intrinsically much less likely to be produced. For instance, the neural basis of the sort of pain typically caused by a stubbing of the toe could not produce a complex visual phenomenology like the one I am having as I stare at my computer—the kind of information it embodies or encodes is of the wrong sort, or not extensive enough perhaps. For now, I am abstracting away from these sorts of obvious limitations on the type of phenomenology the neural bases could produce.

[165] As it almost certainly is—a realistic estimate of the number of distinct phenomenologies is probably much higher, probably even infinite (and perhaps uncountably so).

ability to lead us to expect the evidence. (Recall the earlier discussion of what is required for one hypothesis to receive confirmation/disconfirmation with respect to another one.)

So the central issue at this stage is whether physicalism does lead us to expect the evidence as well as before, or at least better than epiphenomenalism. If objection (H) is mistaken, then I think it does lead us to expect the evidence nearly as well as before, resulting in substantially more confirmation for physicalism over epiphenomenalism than with the original data (since epiphenomenalism does a considerably worse job of leading us to expect the new evidence than the old, as we have seen). The reason is that the physiological constitution of human beings (and in particular, human brains) could not be much different than it is, while preserving organisms' ability to successfully negotiate environmental challenges, at least of the more dramatic sorts. Take the knife cutting example once again. If humans had a much different physiology, it is unlikely that that physiology would get them behaviorally responding in an appropriate way—there is only a narrow range of "hard-wirings" that will produce the requisite behavior (i.e., a withdrawal of the limb) in response to the stimulus. But on physicalism (unlike on epiphenomenalism), there is no contingency in the correlation between phenomenology and physical/functional neural basis. It is a necessary condition (and perhaps a sufficient one as well) for the truth of physicalism that qualitative events metaphysically supervene on physical/functional ones, and so once these physical events are fixed, fully determinate qualitative events are instantiated as a matter of metaphysical necessity. So, if there are very few possible physiological constitutions of human beings that will get them responding to stimuli appropriately, and these physiological constitutions will ensure a particular distinctive phenomenology is instantiated, then there will only be a

few possible distinctive phenomenologies on physicalism, as opposed to a very broad range on epiphenomenalism.[166] (I make the empirical assumption about physicalism, as I have throughout, that all of the possible neural bases that could be produced by typical dangerous stimuli metaphysically necessitate INPs, thus ensuring that physicalism have the consequence that typical dangerous stimuli are guaranteed to be associated with some INP or other.) So, contrast the broad initial way of formulating the evidence with this more fine-grained approach. Epiphenomenalism's probability of expecting each individual correlation on the broad evidence was something in the general vicinity of 1/3, while physicalism's was 1. On the fine-grained evidence, physicalism's is one in whatever small number of possibilities *s* there are, while epiphenomenalism's is one in whatever vast number of possibilities there are (a number certainly more than the roughly 3*s* it would take for the two kinds of evidence to result in arguments that confirm/disconfirm the two hypotheses equally relative to one another).[167]

Indeed, in the coming chapter, I will argue that it is permissible to treat the physical laws as fixed in the context of confirmation this sort of narrow evolutionary argument operates within. (And I will also argue that a closely analogous assumption is permissible when we broaden the argument to include interactionism as a dialectical option.)[168] If I can defend this assumption, we will arrive at an even clearer result. The result is that the physical laws will ensure for all intents and purposes that the only

---

[166] Even if the range of physicalist possibilities is small, there is still the issue that arises if the range is continuous. Because then, for set theoretic reasons, there won't literally be any fewer live physicalist options than epiphenomenalist ones, because the cardinalities of the respective sets will be the same. Unfortunately, I cannot take up this very difficult (but very general) confirmation issue in the present context. For now, I will reluctantly assume that some principled measure can be found the preserves the intuitive difference in the count of possibilities.

[167] Here I assume, as I have done previously in analogous situations, rough indifference in the likelihood that any of the given possibilities is instantiated.

[168] As well as the initial conditions, but this additional assumption is likely to strike people as less controversial.

possible physiological constitution on physicalism is the actual one, and so (by applying the metaphysical supervenience claim discussed above) in turn the only possible determinate phenomenology associated with a given distal stimulus is the actual one. If such an approach is sustainable (and this is contingent on the falsity of objection (H)), it allows us to bring the relationship between physicalism and epiphenomenalism, when evaluated with respect to the different ways of formulating the evidence, into sharper focus. If we treat the physical laws as fixed, then physicalism entails both the broad and the fine-grained evidence, while the probability of the broad evidence on epiphenomenalism is still about 1/3, and the probability of the fine-grained evidence is some much smaller amount. So clearly, in this case epiphenomenalism will suffer considerably more disconfirmation than previously.

Before completing the discussion of variation (3), I should note that in the future I will focus primarily on evolutionary arguments employing maximally fine-grained evidence. Although to this point (in keeping with many historical formulations of the evolutionary argument) I have dealt primarily with versions employing less detailed phenomenological evidence, for the reasons adduced above, the overall argument to be considered in the next chapter will attempt to take into account highly determinate evidence and highly determinate versions of the broad hypotheses (physicalism, epiphenomenalism, etc.). Indeed, the evidence will be more highly determinate than the specific evidence explicitly considered thus far (even in the discussions of this chapter), because information about physiological transitions from stimulus to neural basis of phenomenology, and from neural basis to behavior, will be considered. The motivation for considering this even more highly determinate evidence, as discussed above, are

roughly the desire to keep fine-grainedness of the various kinds of evidence commensurate, and generally to use the most highly determinate evidence possible, in keeping with standard inductive and abductive practice. Hence, the lessons we can learn by paying attention to physicalism vs. epiphenomenalism evolutionary arguments that use very fine-grained evidence are more directly applicable to the overall evolutionary argument, and so it is advantageous for us to begin considering them in more detail.

This brings us to the end of our discussion of alternative formulations of the argument. I hope the reader is now convinced that versions of evolutionary-style arguments against epiphenomenalism of the sort presented earlier which employ variations like the ones embodied in (1) and (2) have roughly the same force as the original. (In fact, not only do they have roughly the same force in the sense of conclusions that are about equally strong, the evidential premises used in the arguments are of similar plausibility to the ones used in the original.) And I also hope the reader is convinced (conditional on the falsity of objection (H)) that versions which use fine-grained phenomenological evidence along with fine-grained distal stimulus evidence (i.e., arguments falling into category (3)) favor physicalism over epiphenomenalism more heavily than do arguments using less fine-grained evidence. Let us proceed, then, to the earlier promised discussion of objection (H)—the objection calling into question the "unbreakable" connection between phenomenology and behavior on physicalism. We will then use that discussion as a springboard to a more inclusive, overall discussion of evolutionary arguments and the mind-body problem.

Chapter 4—The Evidence and the Overall Evolutionary Argument: Physicalism, Epiphenomenalism, and Interactionism

In this final chapter, I will tie together the loose ends of discussions in previous chapters, and explore the overall relevance of the evidence for the mind-body problem. Traditionally, proponents of evolutionary arguments in the spirit of the ones we have been examining—figures such as William James and Herbert Spencer—have believed that the evidence supports both interactionism and physicalism, because interactionism and physicalism appropriate a causal role in the physical world for qualitative events. They have typically believed that the evidence undermines epiphenomenalism, though, because epiphenomenalism does not appropriate such a causal role for qualitative events.

In this chapter, I will argue that they are incorrect. In connection with this project, I will establish two basic claims. They are as follows:

(i) Objection (H) is in fact true—there is no "unbreakable" connection between phenomenology and behavior in a sense to be made more precise shortly, even if physicalism is true.

(ii) Given that objection (H) is in fact true, the evidence we have been considering is useless. Other evidence, along with *a priori* philosophical argument, will settle the issue if anything does. (This other evidence will probably be composed of information relevant to the truth of the thesis of the "causal closure of the physical.")

I will conclude the chapter with some general philosophical morals and discussion of a number of related points.

My route to establishing the two claims will be a bit intricate, so a brief survey is in order. First, in section I, I am going to deliver on a promise from earlier: to defend the claim that, for the purposes of this argument, we can treat the physical laws and initial conditions as fixed, as roughly like necessary truths. Although this assumption is unrealistic according to the most popular theories of laws and causation, I will argue (in ways already hinted at) that it is perfectly appropriate even if those theories are true. (Actually, the claim I will ultimately defend is slightly qualified, but I won't get into the details just yet.) Securing the acceptability of this assumption will make the subsequent discussion proceed much more smoothly and hopefully clearly. In section II, I will more precisely state the different forms that objection (H) can take. I will then isolate the most interesting formulation of objection (H). Section III will show why the anti-epiphenomenalism argument hinges on it. In section IV, I will (at last!) formally introduce interactionism and discuss its dialectical significance (as well as offer a more thorough explanation of why treatment of interactionism has been postponed so long). Then, in section V, I will directly examine the crucial version of objection (H) and conclude (as I said above) that it is successful. Section VI will draw out the implications of the truth of objection (H). Finally, I will conclude with the section on philosophical lessons and morals that come from consideration of the sorts of arguments this dissertation has been concerned with (section VII).

I. Treating the Physical Laws as Fixed

Let me begin by introducing an assumption which, as I discussed above, will make the subsequent presentation proceed more smoothly and clearly. Let us call the assumption the "Narrow Physical Law Fixing Assumption," and define it as follows:

> The Narrow Physical Law Fixing Assumption (NPLFA) := The assumption that the initial physical conditions of our universe and all the laws of nature that actually govern the progression of physical events and the transitions of physical systems can be treated as metaphysically and epistemically necessary truths for the purposes of the anti-epiphenomenalist (and pro-physicalist) evolutionary argument.

Before proceeding to the defense of this assumption, a couple of brief remarks are in order. First of all, it should be noted that there are a number of scientific domains whose events metaphysically and conceptually supervene on the physical.[169] Notable examples include chemistry, geology, and biology (at least where consciousness is excepted from the realm of biology). The distinctive entities in all these domains are uncontroversially composed of purely physical building blocks, and once one is aware of the arrangement of the building blocks (assuming one has a suitably large memory and capacity for conceptual inference), one can see *a priori* that the relevant scientific

[169] Precisely specifying the nature of these supervenience relations is, of course, a difficult task. Consequently, I won't attempt to say anything more about it here. None of the technical details will be relevant for our purposes. Incidentally, I am taking epistemic necessity and conceptual necessity to be the same thing. Thanks to Brian McLaughlin for pointing out the need to clarify the term 'epistemic necessity'.

properties are also instantiated.[170] If NPLFA is justified, then we can justifiably assume that the laws in these scientific domains can also be treated as necessary truths for the purposes of argument.[171]

A second issue that should be dealt with is a potential question the reader may have had about the naming of this assumption—why is it called "narrow"? The 'narrow' refers to the scope of the argument we are employing the assumption in connection with—the argument that pits physicalism against epiphenomenalism only. This is the sort of argument we have been dealing with throughout the work, but later in the chapter, we will have occasion to consider a wider evolutionary argument that brings interactionism into play as a dialectical option. In connection with that argument, our assumption about the fixing of the laws will require some minor qualifications, and hence it will be given a different name.

Let us now continue on to a defense of the assumption. As noted above, there are obvious *prima facie* problems with it. On the metaphysical side, the most popular theories of causation and laws see laws and causal relationships as contingent features of the actual world. According to these theories, in other possible worlds, the properties that play a certain causal role in the actual world play a very different causal role, because in those worlds, the laws are different from what they are in the actual world.[172] But according to NPLFA, we are told to treat the laws as necessary truths, in much the way

---

[170] While many of the properties of these so-called "special sciences" can be realized by multiple arrangements of physical entities (both conceptually and metaphysically), it nonetheless remains the case that there can be no change in them without a change in the physical. It is in this sense that they clearly supervene on the physical.

[171] I assume throughout that the laws of physics are not "gappy"—there are definite predictions made by the physical laws for all physical antecedents (albeit perhaps probabilistic ones in some cases).

[172] The relationship between laws and causality is complicated, and I do not intend my remarks here to imply any substantive commitments beyond the general ones articulated previously.

Shoemaker would have it, meaning that the properties instantiated in the actual world play the same causal roles in all the possible worlds where they are instantiated.[173]

And on the epistemic side, it seems obvious that even if it is metaphysically necessary that the laws be what they are, it is certainly not conceptually necessary that they be what they are. (And it is natural to cash out epistemic necessity in terms of conceptual necessity.) After all, it required (and still does require) elaborate physical experimentation to discover what the laws are; no amount of armchair reflection on our concepts could have alleviated the need for this process!

These are certainly valid concerns, but my approach does not violate the intuitions that underlie them. NPLFA does not claim that the laws *are* metaphysically or epistemically necessary, only that it is acceptable to treat them as such for the purposes of argument.

Odd though the source may seem, the evidence provided by cosmologists is a major reason to treat the physical laws and initial conditions of the universe as a given in the context. Most cosmologists agree that the data their discipline has collected strongly suggests that had the physical laws of the universe been only minutely different, life (indeed, even stars and planets) could not have formed. And without life, the question of evolution and the causal role of mentality in the evolutionary process would not even arise.[174] And moreover, detailed hypotheses (on either mind-body theory) that included

---

[173] Shoemaker, of course, believes all laws hold necessarily. He would believe that if there are laws governing phenomenology that are not ultimately reducible to or metaphysically supervenient upon the physical, then those hold necessarily as well. We are not assuming such an extensive fixing of the laws here. Also, as discussed much earlier, for the committed Shoemakerian about laws, there are analogous assumptions that must be made to achieve the same result that NPLFA is meant to achieve with respect to those who hold that the laws are contingent. (Difficult paraphrases would have to be carried out—in this case, having to do with various properties necessarily being instantiated.) I will not attempt to formulate the assumption in this alternative framework, however, owing to its difficulty.

[174] For a classic and accessible discussion of this evidence, see Leslie (1989).

different physical laws or initial conditions as components would almost certainly conflict with actual observations (or well confirmed theories) of the history of the universe prior to the existence of sentient life on Earth, and so be ruled out anyway.[175] This would be because actual astronomical and geological observations (or well confirmed theories thereof) seem to be part of our background knowledge about the world.

Nevertheless, this is dangerous territory, since considerations like this can easily propel us down the slippery slope epitomized in objection (A) to the original argument—the objection that contended that the evolutionary argument is really based on conceptual considerations rather than empirical ones. How might the considerations propel us down this path? Essentially, by ensuring that no detailed hypothesis is ever allowed to even initially count as an epistemic possibility if it conflicts with our total current evidence.[176] (Recall objection (A)'s contention that because there were determinate forms of the epiphenomenalist hypothesis that led us to expect the evidence we actually find, any preference for the physicalist alternative must have been based on considerations other than those having to do with the evidence on the table.)

---

[175] I admit that a fully detailed or determinate hypothesis/possibility (either physicalist or epiphenomenalist) is a somewhat idealized entity. It specifies the initial conditions of the world, and the subsequent history of the world in its maximal detail (including the relevant stimulus to phenomenology to behavior transitions). (It will also include specifications of the laws themselves, since there are possibilities that agree on the history of the world but disagree on the laws.) In many ways, my detailed hypothesis/possibility is similar to the epistemic "scenario" of Chalmers (2006), albeit with some notable differences. At any rate, the overall probabilities of the general hypotheses—physicalism and epiphenomenalism (and later interactionism)—are a sum of the probabilities of their individual determinate versions.

[176] In somewhat extreme cases, this can lead to ruling out *a priori* from possibility space hypotheses incompatible with one's own existence. Although this may not sound so bad, it can wind up (e.g.) preventing a person from ever gathering evidence in principle that her parents had not used birth control, though there would not be any difficulty in her gathering such evidence about other people (and likewise, none in their gathering such evidence about her). For a view that actually advocates these sorts of extremely counterintuitive inferences, see Sober (2005). There are ways to mitigate the counterintuitive results (if not the counterintuitive process) by claiming that one's prior degrees of belief in various general hypotheses would be affected by the strange *a priori* constraints, but the issues involved are tricky and complicated. I am indebted to David Manley for the birth control example.

Fortunately, we can circumvent worries about sliding down this slippery slope in a principled way. Getting around these worries involves employing a basic strategy we have seen before, in response to the objection previously considered that claimed we were not permitted to use the phenomenology/distal stimulus correlations of other people as part of our argument (objection (F)).

Every confirmation question we consider is set in a context. Sometimes that context is very broad—as when we ask questions about the origin and setup of the universe, or ponder deep skeptical worries about the existence of the external world at all. Other times that context is significantly narrower—as when we try to discern whether an alleged criminal is guilty of the crime he is accused of, or when we weigh the evidence for and against a localized scientific hypothesis (the evidence for a theory of dinosaur extinction, for instance). Typically, in these narrower confirmation contexts it is inappropriate to consider every last epistemically possible scenario believed as part of the relevant *a priori* possibility space.[177] But at the same time neither is the *a priori* possibility space restricted only to detailed hypotheses that already conform to and lead us to expect the total evidence we have at our disposal. There is a stipulated (but not completely arbitrary) distinction typically drawn between evidence and background knowledge, the background knowledge serving to constrain possibility space, while the evidence is used to confirm and disconfirm various general hypotheses (by ruling out specific determinate versions of them). (In broader confirmation contexts, there is little

[177] I use the terminology '*a priori*' loosely here, since if (e.g.) physicalist possibilities are being considered, these may be inconceivable in a familiar sense, and thus arguably not recognized as possible *a priori*, but still somehow recognized as possible. (Also, the fact that some of the fixing of possibility space depends on empirical rather than conceptual considerations seems to count further against the terminology.) The issues lurking here are very deep—I will address them in a bit more detail later, but a full treatment is unfortunately well beyond the scope of this work. For now, I ask that the reader tolerate my terminological looseness.

or no background knowledge employed—little in the case of asking questions about the origins and setup of the universe, arguably none in asking deep skeptical questions about the external world. Consequently, far more of the epistemic possibilities come into play as part of the initial possibility space of confirmation in these instances.)

Unsurprisingly, it is very hard to formulate a general theory of how to draw these distinctions in the various specific cases, and the task has vexed epistemologists, philosophers of science, and philosophers of language to no end. Consequently, I can't hope to say anything new and interesting about the problem in a passing discussion such as this one. What is worth noting is that, even if it is difficult to find a general theory, most people's intuitions yield fairly confident judgments about how to draw the distinction in the specific cases that present themselves (and fairly consistent ones across persons as well). In the evolutionary argument case, although I have not actually conducted a survey, I think it would be fair to conjecture that most people would consider information about the physical laws to be background knowledge in the context, while considering information about (e.g.) what sorts of qualitative events are conjoined with the various stimuli and neural physiological constitutions to be evidence. The intuition behind this common response is strengthened, I think, when we reflect on the fact that any determinate versions of physicalism or epiphenomenalism that have any chance of ultimately being correct are going to agree completely on the progression of the physical world from beginning to end.[178]

[178] The vigilant reader will undoubtedly notice that the claim that we can assume that two hypotheses which make identical correct predictions about the total physical history of the world are qualitatively identical (in the metaphysician's sense of "qualitatively," not the philosopher of mind's) is not strictly entailed by NPLFA, nor does it entail NPLFA. The reason it is not entailed by NPLFA is that even if the laws hold necessarily, if they are indeterministic, this allows for a degree of flexibility in the actual transitions within the physical world. And a major reason why it does not entail NPLFA is similar. If the laws are indeterministic, then presumably even if two fully determinate hypotheses agree on the progression of the

Interestingly, it is possible to offer a further (and more indirect) defense of NPLFA. It can be shown that, given a number of plausible assumptions, NPLFA has no substantive effect on the conclusion of the anti-epiphenomenalist evolutionary argument (so long as fine-grained formulations of evidence about distal stimulus, phenomenology, and physiological transitions are used). In other words, employing NPLFA neither (A) adversely affects the strength of the argument's conclusion, nor (B) positively affects the strength of the argument's conclusion. Given that it has no substantive effect, we can then feel justifiably free to use it at will.

I will not provide this indirect defense, however, for two main reasons. First, the other defense is sufficient to show that the assumption is warranted, and so providing further defenses would be needless. (This is especially so in this case since a long and technical digression would be required.) Second (and to my mind more importantly), later in the chapter we will have to consider a modified version of NPLFA when we examine the broader evolutionary argument (i.e., the one that includes interactionism as a dialectical option). For complicated reasons which I won't address, it is very difficult (perhaps even impossible) to produce an analogous indirect defense of this modified assumption.[179] Since our ultimate concern is the broader evolutionary argument, it seems

---

physical world (where a fully determinate hypothesis specifies in maximal detail the past, present, and future of the world), there is a possibility that they posit different laws which govern the progression of that world. (Assuming they are realists about laws, which we have been assuming from the beginning.) Fortunately, the issues raised by these sorts of concerns are subtle and not really relevant for present purposes (they only introduce the need for tedious, but ultimately insignificant, complication); we can get by largely without dealing with them. We can basically assume that two fully determinate hypotheses agree on the progression of the world from beginning to end if and only if they posit the same laws and initial conditions. On occasion, though, I will address these issues, as relevance necessitates. By the way, there is a further reason why two hypotheses might agree on the physical laws but not on the progression of the physical world: because they posit some other kind of law or causal process that impacts happenings in the physical world besides physical to physical ones. But these will not be physicalist or epiphenomenalist hypotheses.

[179] The issue has to do with differences in prior probabilities among isomorphic determinate hypotheses, but it would require us to lay out a good deal of background in order to go into the matter further.

unnecessary to offer defenses that are not ultimately useful in that context and cannot even be adapted into something useful in it.

II. The Different Versions of Objection (H)

Now that I have defended NPLFA, from here on out I will be employing it in the situation it was designed to be employed in—i.e., where the physicalism vs. epiphenomenalism evolutionary argument is under consideration. It is now time to examine more precisely the different forms that objection (H) can take. When someone claims that there is no "unbreakable" connection between INP (or any determinate kind of INP) and avoidance behavior if physicalism is true, there are several things that the person could mean.[180] Almost trivially, the person could be pointing out that avoidance behaviors do not always follow INPs (even INPs of a specific sort). Imagine a case where one person is pricking me in the finger at the same time that another is standing in front of me with a loaded gun. I have been told that if I don't tolerate the mild irritation and allow my finger to be pricked, I will be shot on the spot. It will not be surprising if I don't engage in avoidance behavior in this situation, even if I am suffering through an INP.

For present purposes, I will ignore these obvious (but superficial) sorts of counterexamples to the loose claim that there is an unbreakable connection between specific kinds of phenomenology and specific kinds of behavior if physicalism is true. I have already acknowledged that the correlations between qualia type and behavior type are not perfectly smooth anyway. Instead, I will focus on more plausible (and more

[180] I will focus here, as I have most often, on INPs and determinate kinds of INPs. The discussion should apply readily to IPPs *mutatis mutandis*.

relevant) challenges to the heart of the crucial thesis. (The heart of the thesis being that necessarily if physicalism is true, there is a strong prevalence of behaviors of a certain sort when a specific type of qualitative event is instantiated, especially when various *ceteris paribus* conditions are met. Recall from earlier that the crucial issue is the relationship between P(*e*/epiphenomenalism) and the P(*e*/physicalism), where *e* is the evidence.[181] Physicalism will be confirmed relative to epiphenomenalism if and only if the latter probability value is greater than the former, and significantly confirmed if and only if it is much greater. In its more relevant forms, then, objection (H) represents a challenge to the claim that the actually observed correlations—the pieces of evidence—are more likely to be the case if physicalism is true than if epiphenomenalism is.)

As I see it, these challenges to the unbreakability can take four different forms, each focusing on a different sense of 'unbreakability'.[182] They are:

(1) The Nomically Accessible Causally Mediated Version (NAC)

(2) The Nomically Inaccessible Causally Mediated Version (NIC)

(3) The Metaphysically Uncausally Mediated Version (MUM)

(4) The Epistemically Uncausally Mediated Version (EUM)

The NAC version of the objection doesn't require the laws to be different from what they are in the actual world (hence the name of the objection). It claims that phenomenology (e.g., a particular kind of sharp pain) could come apart from behavior

[181] In the interest of keeping the discussion as clear as possible, I am omitting reference to background knowledge for the time being. It is not required to appreciate the issues we are dealing with here.
[182] Actually, the third of these four challenges is patently absurd, and so is not more plausible than other attempts not considered. I include it only for clarity's sake—to highlight the differences between it and the fourth challenge.

(e.g., a particular kind of withdrawal of a limb) because it is possible for human organisms to be "engineered" with very different physiology from what they actually have, in such a way that the phenomenology that actually is associated with a particular avoidance behavior could have been associated with some completely different behavior, perhaps even a seeking out. (This is even without any change in the laws of physics themselves.) Thus, the sharp pain that is actually associated with a jerking of the arm away could have instead been (systematically) associated with a seeking out of whatever stimulus had disturbed the arm to begin with. And so, according to NAC, it is wrong to suppose that physicalism is any different from epiphenomenalism—both allow for varied possible correlation relationships between phenomenology and behavior.

The NIC version, on the other hand, does require that the laws be different from what they actually are. This version claims that even if physicalism were true and physiology remained the same, human organisms need not have the same correlations between phenomenology and behavior, because the laws of physics could have been different in such a way that whatever qualitative event is in question could have immediately caused different physical events in the nervous system (from the ones actually caused), ultimately giving rise to a different behavior.

The MUM version, unlike the previous two versions, does not rely on the contingency of the causal connection between phenomenology and behavior to ground the contingency of the correlation between them. MUM instead claims that a qualitative event can come apart from whatever physical event it supervenes upon (i.e., its physical supervenience base), regardless of the breakability between whatever physical event is generated initially by the external stimulus, the neural event that is the supervenience

base of the qualitative event, and subsequent physical events in the nervous system that ultimately lead to behavior. MUM claims that the connection between the supervenience base of a qualitative event and a qualitative event is metaphysically contingent.

Before continuing on to the fourth and final version of objection (H) (which, perhaps unsurprisingly, is the only one I think is ultimately a threat to the narrower evolutionary argument), let me say a few words about these first three versions. There may be a host of problems with NIC, but pointing out one in particular is sufficient to show that NIC is a dead-end worry. NIC requires us to take seriously the possibility that the laws of physics could be different from what they in fact are, and this is counter to NPLFA. Since NPLFA is an assumption that has already been defended (I believe successfully), NIC cannot be entertained.

NAC poses a potentially more interesting challenge, though. There are two broad ways that the physiology of humans could be altered to make NAC plausible—the alternate physiology could be produced by different laws of nature or by the same laws. Obviously the "different laws" road is a non-starter, for the same reason that NIC was. Any claim that requires us to take seriously the possibility that the laws of physics could have been otherwise conflicts with NPLFA.

This leaves us with the "same laws" road. There are in turn two (relevant) broad ways that the physiology could have been altered without the laws being other than they are: there could have been different initial conditions of the universe that resulted in a different subsequent progression of the physical world, or indeterminism in the laws could have allowed a different physiology to be produced along the way from the same initial conditions. Unfortunately, the first of these options still conflicts with NPLFA—

NPLFA assumes not only that the laws are treated as necessary truths, but the initial conditions of the physical universe as well. This leaves us with only the indeterminism option, and unfortunately there are a number of problems with it. A first problem is that some knowledgeable physicists and philosophers of physics do not believe that the laws of physics are in fact indeterministic. Findings in the branch of physics responsible for convincing many of indeterminism, quantum mechanics, are suggestive of indeterminism, but don't straightforwardly entail it. Many have tried to develop non-indeterministic interpretations of the data, most notably Bohmian ones. Though these interpretations are admittedly not the most popular ones available, their ultimate fate is far from sealed. Second, even if the majority is correct and the laws of physics really are indeterministic, it is clear that in normal cases, quantum anomalies have a strong statistical tendency to, loosely speaking, "cancel one another out." This results in an overall system that, at the level of middle-sized physical objects, is overwhelmingly likely to behave in a way virtually indistinguishable from a deterministic one. Since the environment that would have given rise to this radically different physiological construction would have represented a normal case (i.e., the earth in prehistoric times), it is hard to believe that the actual laws would have been responsible for producing it.

Nevertheless, the probabilistic indeterminism associated with the most popular interpretations of quantum mechanics is at least logically compatible with many strange and unexpected occurrences, and so there is no literal inconsistency between NPLFA and the actual laws producing a radically different human physiology (assuming the production of this different physiology was among one of the strange occurrences the laws are compatible with). But given the quasi-miraculous oddity of the statistical

anomalies that would be required for such a progression to take place, there might be reasonable grounds to assign very low prior probabilities to fully determinate hypotheses that included the anomalies. Consequently, even if we were forced to recognize these determinate physicalist options as possible ones in the confirmation space of the argument, we could treat them as insignificant in comparison with the more "well-behaved" physicalist options, where the familiar physiological construction of human beings in the actual world would be preserved, and the overall tendency of (e.g.) specific kinds of INPs to be associated with specific kinds of avoidance behavior in situations where the actual physical laws held would be as well.

In any case, regardless of the prospects of this response, there is a more straightforward reason to reject the indeterministic NAC. The general outlines of the "hard-wiring" of human nervous system physiology are well known, as are the general outlines of the evolutionary development of that physiology. It is highly plausible, then, to suppose that all determinate hypotheses (physicalist and epiphenomenalist alike) that have any chance of ultimately being true will agree on these purely physical transitions, and will entail the observations scientists have actually made in these quarters. But if that is so, it is once again acceptable to treat these physiological constructions and their development as background knowledge—in fact, we could have easily included them in NPLFA along with the initial conditions and physical laws, and done so in a principled way.[183] But if that is so, then it prevents us from taking seriously any version of NAC

[183] One thing that may strike the reader as odd about this claim is that arguments of the sort we are considering are often called (as I have been calling them) "evolutionary arguments." This seems to suggest that the fact that humans evolved, and the circumstances of their evolution, should very straightforwardly count as evidence, not background knowledge. Although I have often framed claims about evolution as evidence, it is possible to view them as background knowledge in a principled way, I think. We can see this most clearly if we examine William James' formulation of the argument from James (1890). There, James says:

that requires us to assume otherwise. Since there are no other versions of NAC available, the overall NAC therefore fails.[184]

Of the three versions of objection (H) that we have considered thus far, that leaves only MUM left. But of all of them, MUM is probably the easiest to refute. MUM is obviously false, because its claims are inconsistent with physicalism (and objection (H) is

---

It is a well-known fact that pleasures are generally associated with beneficial, pains with detrimental, experiences. All the fundamental vital processes illustrate this law. Starvation, suffocation, privation of food, drink and sleep, work when exhausted, burns, wounds, inflammation, the effects of poison, are as disagreeable as filling the hungry stomach, enjoying rest and sleep after fatigue, [and] exercise after rest… are pleasant. Mr. Spencer and others have suggested that these coincidences are due, not to any pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long-run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable… But if pleasures and pains have no efficacy, one does not see… why the most noxious acts… might not give thrills of delight, and the most necessary ones… cause agony.

Notice the quick way in which James (following Spencer) notes the workings of natural selection, and contrast that with the voluminous examples of phenomenology/distal stimulus correlations. It is for him a blatantly obvious fact (thanks to Darwin) that evolution by natural selection has taken place, and he expects all the hypotheses to agree on that as a matter of course. What he finds interesting is rather the correlations between phenomenology and various stimuli in connection with that background fact. These are really his evidential focus. Though we were certainly previously aware of them, the arguments is meant to get us to appreciate their significance in the light of the background considerations.

For those not convinced (e.g., those who believe this move can be shifted around just as easily—that the argument wants us to appreciate the evidence of natural selection in the light of our background knowledge of the correlations), please stay tuned for the concluding section of the chapter. There, I argue that when a mature version of the argument is examined and very fine-grained evidence is taken into account (including evidence about physiological transitions in the nervous system), evolutionary evidence plays no role in driving us to the conclusion. In fact, even evidence about the survival of presently living organisms plays no part.

In any case, another thing that should be noted is that I will ultimately be arguing in favor of objection (H) in one of its forms, so any mistakes I make along the way in rejecting other versions can only help my ultimate conclusion when corrected, by preserving the plausibility of these other versions as well.

[184] Incidentally, there is a version of objection (H) that is an amalgam of both NIC and NAC. This version claims (like NIC) that the physiology of humans is the same and (like NAC) that the laws of physics are the same as in the actual world, but asserts that the connection between qualitative event and behavior comes apart because indeterminism leaves open the possibility that qualitative events (e.g., sharp pains) luckily cause mostly *prima facie* unexpected behaviors (e.g., seekings out), in virtue of immediately causing *prima facie* unexpected physical events in the nervous system. My response to this sort of objection should be fairly clear from the discussion above. First, it presupposes indeterminism in the laws, which, while a popular view, has not been firmly established. Second, even if indeterminism is true, it is not clear that the indeterminism quantum mechanics allows for is robust enough to affect physiological transitions in the brain of an organism in a noticeable way. For those still holding out hope for the objection, although one cannot make a response analogous to the background knowledge one for NAC above (since it seems that information about physiological transitions in the brain and the like is evidence, not background knowledge), one can still make the analogous response about prior probabilities—fully determinate physicalist hypotheses that called for systematically unusual physical transitions would be assigned a low prior probability, and thus not count for much in the overall appraisal of what physicalism as a general hypothesis would lead us to expect.

an objection that is meant to apply to physicalism). Physicalism claims that phenomenology metaphysically supervenes on its physical neural base. It is metaphysically necessary that once the appropriate physical neural event transpires, the qualitative event does as well. So claiming both that physicalism is true and that a physical neural event *P* (actually associated with qualitative event *Q*) can be instantiated in another (physicalist) possible world without the accompanying *Q* is inconsistent. But this is just what MUM is claiming can occur.[185]

Now that we have seen the pitfalls of the first three versions of objection (H), let us continue on to the fourth—EUM.[186] Unlike the metaphysical causally unmediated version, this epistemic version is not obviously false. In fact, at least *prima facie* it seems true. It claims that regardless of the metaphysical relationship between qualitative events and the physical events that are their bases, it is clear that there is no epistemic supervenience (i.e., something like logical or conceptual supervenience) of phenomenology on the physical. This is, of course, a familiar mantra from the last few decades of debate on the mind-body problem, and the same basic intuition behind the idea has been expressed in countless different ways. Although the intuition has been challenged on occasion, the challenges have been rare and generally not compelling. The main concern is not so much whether EUM is true, but rather, the significance of its truth for the arguments we are considering here. More specifically, the crucial question is whether the truth of EUM should affect our appraisal of the probability of the evidence

[185] Functionalists, of course, are typically willing to deny $\Box(Q \rightarrow P)$, because they believe that qualia can be realized by different physical events (or something similar). But all that is at issue here is the converse of this claim, $\Box(P \rightarrow Q)$. I am not aware of anyone who would deny this claim (assuming we are restricting our attention to physicalist worlds), at least where qualitative events and physical events are being described in an appropriate way.

[186] Let me remind the reader once again that I am ultimately arguing that objection (H) is a success. Hence, any mistakes I have made in challenging the first three versions of objection (H) can only strengthen my conclusion, by preserving the plausibility of versions of the objection that I have dismissed.

given physicalism, and by extension the power of the evidence to confirm physicalism over against epiphenomenalism. It is to this issue that I will turn in the next section.

III. Objection (H) and the Anti-Epiphenomenalist Argument

We have now come to the point where it is clear that the only potential threat to the anti-epiphenomenalism evolutionary argument is objection (H) (at least with any immediate chance of working), and the only interpretation of objection (H) that could constitute this threat is EUM. Seeing why objection (H) could potentially constitute a threat is not hard, but it is worth taking a few moments to clarify the nature of the looming difficulty. (Some of this is basically review from earlier, but stated a bit more precisely in light of the intervening work on objection (H).)

Basically, there are two general possible answers to the question of EUM's impact on our judgments about the relevant probability. Either it does warrant us in thinking that P(*e*/physicalism) is significantly lower than the argument claims, or it does not. One might (at least *prima facie*) think it does not, since the possible variation in question is merely epistemic if physicalism is true; there are no metaphysical consequences of it. (Just because I can conceive of a particular physical neural event actually accompanied by a specific painful qualitative event accompanied by some different sort of qualitative event, or no qualitative event at all, has no metaphysical implication according to physicalism.) If the truth of EUM is ultimately insignificant in evaluating P(*e*/physicalism) (and thus it does not convince us that the relevant probability is lower than the argument claims), then this is ultimately going to result in the failure of

objection (H) to convince us that physicalism does no better job than epiphenomenalism of leading us to expect the evidence we find. Consider the following diagram:[187]

EPIPHENOMENALISM

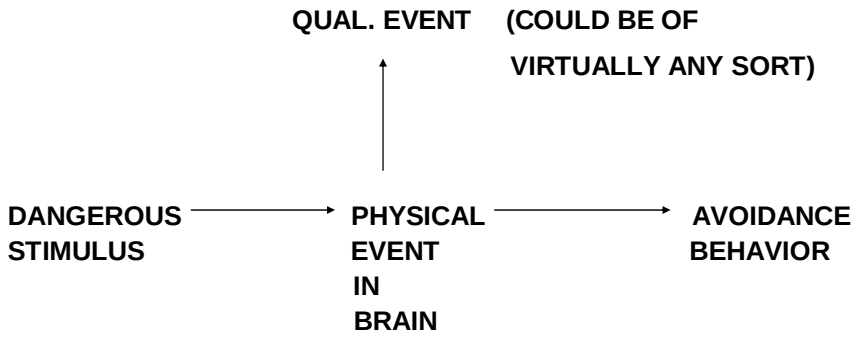

FIGURE 1

As we have been through on many occasions, we know that if epiphenomenalism is true, there could have been tremendous variation in the sort of qualitative event that accompanied some physical event in the brain produced by a specified dangerous stimulus. And moreover, this "could" is a "could" of both metaphysical and epistemic possibility. Obviously, I could conceive of it being different, but also, according to epiphenomenalism, the laws of nature could have been different in a metaphysical sense, and some of those different laws would have produced markedly different qualitative events. Contrast this with the following diagram, though:

[187] Note that the arrows represent causal connections, presumably governed by laws of nature. By the way, I do always assume the bridge laws from physical/functional to qualitative are deterministic. To do otherwise would complicate matters beyond what is tolerable, and I'd speculate have little affect on the overall force of the argument. What I'm not assuming are deterministic are the physical laws governing the causal transitions among physical entities.

PHYSICALISM

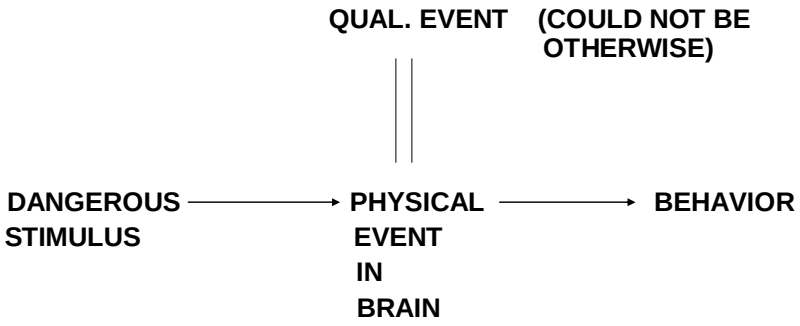

FIGURE 2

If physicalism is true on the other hand, and EUM's epistemic breakability is merely one of conceptual fancy, with no implications for metaphysics or confirmation, then physicalism will be able to rely on the metaphysical supervenience of qualitative event on physical neural event to secure a high P(*e*/physicalism), in comparison with a low P(*e*/epiphenomenalism).

But, if EUM's epistemic breakability does have implications for confirmation, then the physicalist picture will instead look like this:

PHYSICALISM ALTERNATIVE

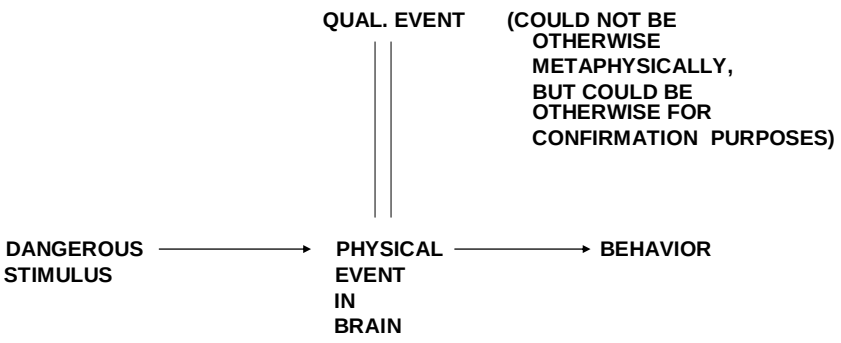

FIGURE 3

If EUM's epistemic breakability is allowed to have implications for confirmation (even if not for metaphysics), then for every determinate epiphenomenalist hypothesis there will wind up being a corresponding determinate physicalist one. Since there would seem to be no *a priori* reason to treat the prior probability of any determinate physicalist hypothesis any differently from its epiphenomenalist isomorph, the result would be that P(*e*/physicalism) would equal P(*e*/epiphenomenalism), spelling doom for the anti-epiphenomenalist argument. (For each live epiphenomenalist possibility, there would be a corresponding physicalist one. When the evidence was gathered and individual determinate versions of physicalism and epiphenomenalism were ruled out because they conflicted with the evidence, an equal number of physicalist and epiphenomenalist

possibilities would remain. Thus, after renormalization, relative to one another both would occupy the same-sized "slice of the overall pie" as before.)[188]

So, now that we can hopefully see more clearly what it is at stake, let us turn to a consideration of interactionism. Once interactionism has been discussed, we will finally be in a position to examine the overall evolutionary argument, as well as finish our treatment of the physicalist vs. epiphenomenalist one, in connection with that examination.

IV. Interactionism: A Formal Introduction and Discussion

Definitions and Background

As noted above, traditional proponents of evolutionary arguments in the mind-body sphere did not typically restrict their attention to arguments that took only epiphenomenalism and physicalism seriously. And rightly so—the restriction is somewhat artificial, since it results in an important dialectical option being left off the table: interactionism.[189] For our purposes, let us understand interactionism as follows:

> Interactionism := Dualism and at least some qualitative events are causally efficacious with respect to physical events

[188] I assume here that the number of determinate hypotheses on both the physicalist and epiphenomenalist side is finite. This assumption may be unrealistic; in reality, there may be an infinite number of such possibilities, perhaps uncountably many. This may cause mathematical difficulties, since the intuitive "size difference" between sets that holds in the finite case may fail to hold. These sorts of difficulties are unfortunately well beyond the scope of what I am able to discuss here, and I must rely on there being some measure that preserves the intuitive differences in a principled way in the infinite cases, without actually specifying how this would work.

[189] There were, of course, reasons for our exclusion of it prior to now. Some of these reasons have already been discussed, and I will have more to say subsequently.

As such, interactionism is a fairly straightforward position (we have already discussed each of its components in connection with the other general views), and may even be the dominant mind-body theory among the mass of humanity (though probably not among philosophers).

In the past, most philosophers who employed evolutionary arguments that took interactionism seriously along with physicalism and epiphenomenalism (henceforth, these will be called "broad evolutionary arguments") came to a characteristic conclusion about the fate of interactionism. Call this the "standard conclusion":

> The Standard Conclusion := When we take into account the evidence, only epiphenomenalism is disconfirmed. Both physicalism and interactionism are confirmed, because both allow for qualitative events to be causally efficacious in the physical world.

Up till this chapter, there has been little reason for us to include interactionism in our discussion. The basic move of the argument could be appreciated without involving interactionism, and all of the objections we've examined have either been specific to epiphenomenalism or raised general issues about the argument strategy employed. Discussing interactionism earlier would have only been a distraction and required the addition of straightforward but tedious applications (to the interactionist case) of the generic responses. But since objection (H) targets the central dialectical move, bringing interactionism in at this juncture is a must. In large part, this is because in the past

philosophers have been apt to make mistakes concerning the relevance of the evidence when interactionism is introduced.

My specific aim in the interactionism discussion is twofold. First, to convince the reader that if objection (H) is false—i.e., (building on our previous results) that EUM has no relevance for confirmation purposes—then the standard conclusion is false, or at the very least far too simplistically stated. Only physicalism gets support from the evidence, if the evidence is what it is typically taken to be. Interactionism does not. But, if the evidence is not what it is typically taken to be (more on this later), just the opposite is true: interactionism gets support, and physicalism does not. And second (as I have already mentioned), to show that in fact objection (H) is true (i.e., that the EUM version has implications for confirmation), and as a result the evidence about phenomenological correlations is utterly useless. Only other evidence and *a priori* philosophical argument will settle the issue if anything does.

Consequently, we will see that traditional defenders of broad evolutionary arguments were wrong in two separate respects—they were wrong in how they treated interactionism given the central assumption of their arguments, and they were wrong in making the central assumption in the first place.

The second of these main aims will have to wait for a direct examination of EUM in the next section, but I will tackle the first in this section. Before doing so, though, we will have to brave a brief digression on extending NPLFA to broad evolutionary arguments.

NPLFA and Broad Evolutionary Arguments

We cannot simply extend NPLFA to broad evolutionary arguments because interactionism implies that the progression of the physical world will not be governed solely by physical laws. Because interactionism allows for qualitative events to have causal efficacy with respect to physical events, there is no guarantee that those qualitative events will cause the same physical events, in the same ways, as the physical neural underpinnings of the qualitative events would if they occurred alone. (Whereas we could assume this when only physicalism and epiphenomenalism were in play, since neither of them allowed any role for anything but physical events to play a causal role in the physical world.)

To sharpen the discussion a bit, let us introduce several new terms. The first, 'The Predictability Thesis', and the second, 'physical progression profile', will be used in the definition of the third. We will understand the Predictability Thesis as follows:

> The Predictability Thesis := For all closed systems *s* outside a brain and all closed systems *t* inside a brain in all worlds *w'* with the same laws as the actual world and no properties alien to the actual world instantiated, if the physical initial conditions of *s* and *t* are identical, then the physical progression profiles of *s* and *t* will also be identical.[190]

And a physical progression profile will be understood as follows:

---

[190] The addition of the caveat about all the worlds *w'* is to prevent the thesis from being trivially true, in case in actual fact there are no such matching systems. I include the caveat about alien properties to preclude the possibility of an intuitive counterexample to the principle which could occur if the actual world includes fundamental psychophysical laws but no fundamental qualitative properties to enter into them.

> Physical Progression Profile of System *p* := If the laws of nature are deterministic in the possible world *w* where *p* occurs, then the overall fully detailed dynamical history of the physical entities in *p*. If the laws are indeterministic, then the overall probabilistic distribution of possible fully detailed dynamical histories of the physical entities in *p*, ruling out any changes in the laws or interference in the system.[191]

A couple of quick points about the definition of 'physical progression profile'. First, it should be noted that I am understanding a dynamical history to exclude facts about causation and governance by laws. It is meant to include only things like the motion of the physical entities in question and their purely structural relationships with one another. Second, strictly speaking, the division into the two conditionals is unnecessary. If the consequent of the second conditional were the entire definition, this would suffice. The first conditional would then be redundant, and deterministic systems would simply be a special case; this is because the entities in deterministic systems would have only one possible dynamical history, and it would have probability 1. I include the separate conditionals here just for clarity's sake.

We are now in a position to define a term that we will have occasion to use often hereafter, 'uniform physicalism':

> Uniform Physicalism := Physicalism and the Predictability Thesis

[191] I assume that the only way for laws to be indeterministic is to be probabilistic.

Intuitively, the Predictability Thesis states that the transitions in physical systems inside the brain do not behave in any way differently from those outside the brain. There are no special laws that govern these transitions (at least not that produce results different from what the other laws would tend to produce) or exceptions to what the existing physical laws would lead us to expect. And so uniform physicalism claims this, in addition to claiming physicalism. (Most physicalists would undoubtedly be uniform physicalists.) If we knew that the Predictability Thesis were in fact true, we could simply extend NPLFA to broad evolutionary arguments and treat the progression of the physical world as background knowledge common to all determinate hypotheses—interactionist as well as epiphenomenalist and physicalist. But it is of course a controversial matter whether the Predictability Thesis is true—although few (perhaps no) physicalists or epiphenomenalists dispute it, many interactionists do, claiming that qualitative events causally influence happenings in the physical world in a way that "physics outside the brain" would not lead us to expect. Consequently, (as we will see in more detail below) assuming it would be begging the question against interactionism and stacking the evidential deck against it.

However, there is something in the neighborhood of NPLFA that interactionists can surely accept. Since their only potential issue is with the progression of the physical world where (qualia producing) brains are concerned, they can surely accept as background knowledge the initial conditions of the world, the physical laws governing physical systems that do not include brains, and presumably (in keeping with the extension of NPLFA discussed earlier) even the progression of the physical world up to

the time that sentient beings first appeared.[192] Call this modified assumption the "broad physical law fixing assumption," or 'BPLFA' for short.

Some Final Preliminary Material

Now, let us return to our primary remaining task for the section—showing that if objection (H) is false, then traditional broad evolutionary arguments have tended to mishandle interactionism. There are two pieces of the puzzle we still need to put in place in order to do this properly—formulate a paradigmatic and fairly precise broad evolutionary argument, and briefly discuss the prospects of the other versions of objection (H) (NIC, NAC, etc.) that we dismissed in connection with the narrower evolutionary argument.

Let me begin by providing a paradigmatic formulation of a broad evolutionary argument. This will allow us to examine objection (H)'s significance in the context more precisely. We have already seen the argument's basic contentions about physicalism and epiphenomenalism's ability to explain the evidence. (I.e., the probability of that evidence given the hypotheses). But now that we are looking at both a broader and more precise version of the argument than we formulated in the opening chapter, a review is in order. The argument claims that physicalism does a good job of leading us to expect the evidence we find (i.e., evidence about correlations of distal stimulus with phenomenology, as well as—in keeping with our principled use of the most determinate evidence available—the physiological transitions leading from stimulus to neural basis of phenomenology to behavior), because physicalism allows qualitative events to play a

[192] I assume here the falsity of extreme views like panpsychist interactionism. I take it that this is not an overly ambitious assumption.

causal role in behavior. It claims that epiphenomenalism does not, on the other hand. The alleged reason is that epiphenomenalism does not allow for qualitative events to play a causal role in behavior.

According to these traditional broad evolutionary arguments, interactionism also does a good job of leading us to expect the evidence. This is because interactionism, like physicalism, allows qualitative events to play a causal role in behavior. So because physicalism and interactionism both allegedly do a good job of leading us to expect the evidence we actually find, and epiphenomenalism does not, epiphenomenalism is disconfirmed and the other options confirmed.

It will be useful for us to express these results a bit more formally, and to construct a model that will help us visualize the confirmation claims being made. According to confirmation theory, a piece of evidence confirms a hypothesis overall (as opposed to relative to another hypothesis, which is the sort of confirmation we were concerned with when examining the narrower argument) if and only if the hypothesis is more likely on the evidence (plus background assumptions) than is the hypothesis on the lack of the evidence. In turn, this is true if and only if the evidence is more likely on the hypothesis than on the hypothesis's negation.[193] To put it formally: P(h/e and k) > P(h/k) iff P(e and k/h) > P(e and k/~h), where e is the evidence, k is the background knowledge, and h is the hypothesis. Since the general hypotheses of physicalism, epiphenomenalism, and interactionism are mutually exclusive and jointly exhaustive, we can represent the

[193] Again, there are numerous qualifications that must be made to this general contention about confirmation, but all are technical and none are particularly relevant for our purposes.

overall probability of physicalism (e.g.) once the evidence has been taken into account as follows, using an elementary application of Bayes' Theorem:[194]

$$P(\text{physicalism}/e) = \frac{P(\text{physicalism}) * P(e/\text{physicalism})}{P(e)}$$

P(*e*), in turn, is calculated as follows:

$$P(e) = P(\text{physicalism}) * P(e/\text{physicalism}) + P(\text{epiphenomenalism}) * P(e/\text{epiphenomenalism}) + P(\text{interactionism}) * P(e/\text{interactionism})$$

Here, P(physicalism) is the probability that physicalism is true prior to consideration of the evidence we are now examining, P(*e*/physicalism) is the probability of our receiving this evidence conditional on physicalism (convincing us that this value is high is a key aim of both broad and narrower evolutionary arguments), and P(*e*) is the overall probability that we would receive this evidence prior to our receiving it.

I will not go through each of the three cases, but hopefully the reader can see that overall probabilities for each of the three general hypotheses can be calculated in a similar way. Where 'physicalism' appears in the numerator in the above calculation, the name of the respective hypothesis is substituted. The denominator remains the same, since P(*e*) is unaffected by the specific hypothesis under scrutiny at the moment. It is not hard to see that a hypothesis will be confirmed (i.e., have its overall probability raised by

[194] For clarity's sake, I will once again omit reference to background knowledge.

consideration of the evidence) iff P(*e*/*h*) > P(*e*). (The claim about what confirmation consists of several paragraphs above entails this claim as well.)

In order to visualize this confirmation process in action, imagine a giant circle. The circle represents the overall space of hypotheses—it is divided into three equally sized regions, each allocated to one of the respective general views (epiphenomenalism, interactionism, and physicalism). Each region is itself divided into smaller equally sized and uniformly shaped regions, one for each fully determinate version of each general hypothesis (specifying the entire alleged history of the world in all its detail, including the laws and operative causal processes).[195] Rising vertically from each small region (into the 3rd dimension) is a volume, representing the overall probability that the determinate hypothesis represented by that small region is in fact actual. (The volumes will sum to 1, since the probability is 1 that the disjunction of all the individual determinate hypotheses is true—after all, they are mutually exclusive and jointly exhaustive, so exactly one of them must be actual. The key is figuring out which one!)[196] Each individual volume will be determined by two variables—the intrinsic (*a priori*) probability of that particular determinate hypothesis and its fit with evidence thus far encountered (i.e., background knowledge).[197] Since it is a difficult question unto itself

---

[195] I assume, as previously, that there a finite number of these regions, and also that there are no duplicate hypotheses. The former assumption is probably unrealistic, since there seem to be (uncountably) infinite variations on each of the general hypotheses. The model is purely for heuristic purposes, though—I admit that a fully realistic portrayal would require us to tackle the mathematical issues surrounding these infinite cases. Unfortunately, though, this is well beyond the scope of the present project.

[196] The probabilities represented by the volumes are of course epistemic, and are relative to an individual's epistemic position (i.e., what evidence has been thus far encountered). Throughout our discussion, we are assuming the individual involved is a generic person in possession of all of the relevant empirical evidence that human beings have gathered.

[197] Readers will no doubt recognize that I am making a substantive assumption here—that some version of subjective Bayesianism is false. Subjective Bayesians think (roughly) that there are no normative constraints on what I have called intrinsic probabilities, aside from logical consistency with probability theory and perhaps some technical constraints that are irrelevant for our purposes. (I.e., a person can feel

what in turn determines the values of these variables in the specific cases, a few brief remarks will have to suffice for present purposes. The intrinsic probability of a hypothesis (i.e., its probability before any evidence is received) is largely—or perhaps entirely—going to be a matter of simplicity. The nature of simplicity itself is too difficult to pin down precisely in offhand remarks; indeed, the task has vexed philosophers of science for ages. The idea is that it will have something to do with the number of entities postulated, the sorts of entities postulated, the regularity of their behavior, etc. The simpler a hypothesis is, the higher the probability *ceteris paribus*. Fit with evidence already encountered may not admit of degrees—the answer may simply be "is compatible" or "is not compatible."[198] Determinate hypotheses that are incompatible with evidence already encountered will have their volumes shrink to 0, as they are no longer ways actuality could turn out. The total of the overall volume that they previously occupied is then distributed to other hypotheses (ones that have not been definitively ruled out), according to some renormalization rule. Although specifying the nature of this rule is extremely difficult in the general case, I will venture the following working view of the renormalization rule for our purposes:

> When a fully determinate hypothesis *h'* has volume *v* at epistemic position *x* and volume 0 at epistemic position *x*+1 (i.e., after the next piece of evidence is considered), volume *v* will be distributed to all the determinate *h* with volume ≠ 0 at *x*+1 proportional to their volume at *x*.

---

free to assign whatever intrinsic probabilities she likes to the individual small regions, so long as these are logically consistent with probability theory.)

[198] For present purposes, I must once again introduce a simplifying assumption: that any time we gather evidence, we know that evidence with absolute certainty. There are definitely confirmation approaches available to deal with evidence not known with such certainty, most notably "Jeffrey Conditionalization," which I alluded to in an earlier chapter.

When new evidence is encountered, then, it changes the overall distribution of volumes in the landscape. The new evidence is incompatible with some determinate hypotheses, and they in turn will lose all of their volumes, never to regain them again. The volume that they used to have is then siphoned off to other hypotheses—if, e.g., hypothesis #1 and hypothesis #2 are both compatible with this new evidence, and hypothesis #1 had probability *a* previously and hypothesis #2 had probability 2*a* previously, then hypothesis #2 will receive twice as much of this siphoned off volume as hypothesis #1. Over time (if all goes as it typically does when gathering evidence), as more evidence is gathered, more and more of the volume will be concentrated in fewer and fewer of the small regions.

As we go along in the process, the probability that each general hypothesis is true will always be the sum of the volumes of the smaller regions that represent its determinate versions. So the overall volume (i.e., probability) of interactionism, for instance, will be a sum of the volumes of all the sub-regions of the overall interactionism region of the circle.[199]

So, applying this general picture to the argument at hand, traditional proponents of broad evolutionary arguments, in defending the standard conclusion, have contended that after the stimulus-phenomenology correlation and evolutionary evidence is taken into account (plus whatever other determinate evidence it is appropriate to take into account, such as physiological transition evidence), the proportion of the overall volume occupied

[199] Although many of the details have been adapted, I owe the spirit of this general visual aid to Meacham (unpublished).

by interactionist and physicalist regions increases, while the proportion of the overall volume occupied by epiphenomenalist regions decreases.

Now that we have a more precise feel for the overall broad evolutionary argument, on to our next task—discussing the disgraced versions of objection (H). Since objection (H) really has to do with physicalism's ability to lead us to expect the evidence, not interactionism's, there is no direct reason why introducing interactionism into the mix would have any effect. The main reason why interactionism might have an effect is instead indirect—because we adopted something weaker than NPLFA (i.e., BPLFA)[200], there is the possibility that we undercut some of the previous arguments against NIC, NAC, and/or MUM.

It is clear that nothing about the weakening of NPLFA affects our previous arguments against MUM. MUM is simply plainly false because it contradicts a clear implication of physicalism—that qualitative events metaphysically supervene on physical ones. (Similarly, though it is not at issue here, nothing about EUM will be affected by introducing interactionism and weakening NPLFA, since our evaluation of EUM was unrelated to NPLFA.)

It is a little less clear with NAC. (Recall that NAC claims humans could be engineered differently, resulting in different correlations.) One of the central contentions offered previously against NAC was that it conflicted with background knowledge about the physiological development of human beings. But if we are only taking for granted the progression of the physical world up to the time that sentient beings first appeared, it seems there is plenty of room for the laws to produce human beings with different

[200] At least weaker than NPLFA where NPLFA is extended to include the previous physical history of the world and not just the laws and initial conditions.

physiological constructions from the actual one. Two things to consider, though—first, presumably whatever changes lead to the altered physiology will be gradual. (Physicalist hypotheses that predict sudden and dramatic changes will either be impossible, or will require such miraculously improbable physical anomalies that they will receive extremely low prior probabilities, and so—as we have already seen—have little influence on what physicalism is overall likely to produce.) In order to be of any real significance to the argument, whatever spurred the change would have to be something that interactionism would lead us to expect better than physicalism or epiphenomenalism would, or vice-versa. (This is because, once we took account of the fact that these transitions had not occurred and ruled out the determinate hypotheses that predicted them, the only way this would substantively change the overall probabilities of the general hypotheses—compared with what they would be if we merely treated the transitions as background knowledge—is if one or two of the general hypotheses has a disproportionate amount of its overall "stake" in the pie invested in determinate hypotheses that predict these transitions. In other words, to use our visual aid, if a disproportionately large amount of the total volume occupied by the regions of one or two of the general hypotheses is disproportionately allocated to sub-regions that predict these transitions.) Seemingly, the only plausible ways this could be the case is if (A) the transitions were brought about by a qualitative event in a pre-human organism, producing some effect that is counter to what we would expect the qualitative event's physical base to cause (hurting interactionism when they were discovered not to actually be the case), or (B) the transitions were brought about in keeping with The Predictability Thesis, but counter to what interactionism would predict (because interactionism would predict that qualitative

events would interfere in this process in some way, thus hurting epiphenomenalism and uniform physicalism—presumably along with physicalism generally[201]—when they were discovered not to be the case).

But, ultimately it's hard to convince oneself that there would be these striking differences in the hypotheses' ability to predict these physiological changes. What qualitative event could be pervasive enough, in a large enough sampling of determinate interactionist hypotheses, to systematically either produce such changes (contrary to what the Predictability Thesis would suggest) or impede them? Consequently, we can conclude either that NAC is still false for broad evolutionary arguments (just as for narrower ones), or that NAC is irrelevant, because even if it is a correct way of cashing out objection (H), it nevertheless has no substantive bearing on a broad evolutionary argument.[202]

That leaves us with only NIC remaining. Our only argument previously against NIC—a perfectly good one in the context of the narrow anti-epiphenomenalist argument—was that it conflicted with NPLFA. But now that we are considering broad evolutionary arguments, we are no longer taking for granted as background knowledge the laws governing the behavior of physical entities inside the brain. (And these are the kinds of laws NIC is considering.) Consequently, NIC is unscathed by previous arguments. NIC also leads us into some of the deeper issues surrounding broad

---

[201] More on the relationship between uniform physicalism and physicalism generally coming up.

[202] It bears repeating also that I am ultimately arguing that objection (H) is correct—i.e., that objection (H) destroys the force of both the broad and narrow evolutionary arguments. Though I have only argued that the EUM form of the objection is plausible (though the NIC form clearly gains some plausibility when NPLFA is relaxed), I would certainly welcome other versions of objection (H) that had promise. I believe NAC does not, but if it did, this would not be a problem. (There is, of course, the issue of the other part of my appraisal of broad evolutionary arguments—that they have traditionally mishandled interactionism even if objection (H) is mistaken. But notice that this claim is conditional on objection (H) being mistaken, and so on all the versions of objection (H)—including NAC—being false anyway. So its plausibility is undisturbed by any mistakes I may make in arguing that NAC is false or irrelevant.)

evolutionary arguments, and so considering it provides us with an opportunity to delve directly into those issues. Although we will ultimately see that NIC fails to undermine physicalism's ability to account for the evidence, the best way to do this is by examining the broad argument and then applying the lessons to NIC afterward. For the moment, then, let us continue to assume that the only way that objection (H) could be threatening is in its EUM form.

In this subsection, we set out to accomplish two tasks. First, to formulate a reasonably precise version of a broad evolutionary argument, and second, to ensure that our earlier dismissals of the three non-EUM versions of objection (H) could be maintained in the context of the broad evolutionary argument, since our earlier dismissals made use of assumptions specific to the treatment of the narrower argument. I hope that all of these tasks have been accomplished successfully, with the exception of the postponed discussion of NIC.

Broad Evolutionary Arguments and Objection (H)

We are now finally in a position to appreciate how broad evolutionary arguments have mishandled interactionism (mishandled it assuming objection (H) is misguided). Let us assume then, for the (temporary) purposes of argument, that objection (H) fails (in its EUM form), and that physicalism does a very good job of leading us to expect the evidence we actually find. Because there is a tight metaphysical connection between the supervenience base of a qualitative event and the qualitative event itself according to physicalism, there is very little variation we will expect to find in the sorts of stimuli that

cause specific qualitative events or the sorts of behaviors those qualitative events are associated with if physicalism is true.

Consider now a particular kind of sharp pain caused by our paradigmatic cutting of the arm. Suppose also for the moment that the Predictability Thesis is true—i.e., that physical entities in the brain behave in the same ways as physical entities outside the brain. This sharp pain is going to be closely correlated with a specific environmental cause and characteristic behavior pattern. Call the causal profile that this sharp pain is typically tied up in ϕ. If ϕ is the dominant causal profile for this sort of sharp pain to be a part of in worlds with the actual laws and basic background conditions of our world, there are only two ways the sharp pain could be part of some other causal profile (excepting cases where *ceteris paribus* conditions fail, as when someone has an even worse pain in another part of the body, or has an unusual belief or desire that motivate her to act in strange ways in the circumstances, etc.). These two ways are: (A) if the physical laws were different, or (B) if unusual and perhaps systematic indeterministic anomalies occurred in the physical laws that allowed for the alternate causal profiles.

Now, think about any determinate version of physicalism where either (A) or (B) was the case. If (A) were the case, and the determinate hypothesis was incompatible with background observations about the history of the physical world (in keeping with BPLFA), then we needn't worry about it. It's overall probability (it's overall volume, to use our visual metaphor) will go to 0 as soon as this background information is accounted for, and it will not be a contributor to physicalism's overall probability at the end of the day. But if it is compatible with background observations about the history of the physical world, then it is overwhelmingly likely that it implies the falsity of the

Predictability Thesis.[203] But notice what this will imply about the laws of physics in this determinate version of the physicalist hypothesis—it will imply that they are weirdly gerrymandered. Special physical laws will apply inside the brain, while other physical laws will apply outside it, even when the systems being governed are otherwise identical. Presumably, this will complicate the hypothesis, and make its intrinsic probability low in comparison with other, more regular, hypotheses. Similar considerations are in play when we consider hypotheses that fall under (B). If the hypotheses continually predict freakish out of the ordinary transitions amongst the physical entities, they will suffer from a low intrinsic probability as a result. (Even if the transitions they predict are no more freakish than any other transitions compatible with the probabilistic laws, they will represent such a tiny fraction of overall hypotheses that their contribution to the overall intrinsic probability of the general hypothesis will be negligible.)

There is a moral to be gleaned from all this. Detailed physicalist hypotheses that represent something other than the Predictability Thesis's truth are responsible for only a relatively small portion of the overall intrinsic probability of physicalism. Because their intrinsic probabilities tend to be low owing to their gerrymandered laws or freakish anomalies, the more regular versions of physicalism will tend to be the bulwark behind physicalism's overall intrinsic volume. Accordingly, if physiological observations are made that fit well with the Predictability Thesis, the vast majority of physicalism's volume prior to those observations being made will be preserved, since little of that volume will presumably be tied up in determinate hypotheses that conflict with it. If

[203] I say "overwhelmingly likely" rather than "certain" because the hypothesis could imply the correct actual history of the physical world by implying that a large number of freakish probabilistic anomalies had occurred. The hypotheses that fit into this category and slip through the cracks of my discussion of (A) will be covered under the discussion of (B).

other general hypotheses have proportionately more of their volumes tied up in such determinate hypotheses (or have other difficulties conforming to actual observations), this well spell good news for physicalism and bad news for them.

But of course there's a flip side to this boon for physicalism. If physiological observations fit in poorly with the Predictability Thesis, then physicalism will lose much of its volume, because so much of it is tied up in determinate hypotheses that affirm the Predictability Thesis. If other general hypotheses do not have as many of their eggs in the Predictability Thesis basket, they will benefit at physicalism's expense.

Let us turn our attention briefly to epiphenomenalism once again. Because epiphenomenalism allows only physical entities, obeying physical laws, to play a causal role in the physical world, determinate epiphenomenalist hypotheses that violated the Predictability Thesis would similarly constitute a relatively small portion of epiphenomenalism's overall intrinsic volume, and presumably also volume after background knowledge is taken into account. And so epiphenomenalism, similar to physicalism, would seemingly benefit from physiological observations that fit in well with the Predictability Thesis and be undermined by observations that fit in poorly with it.

But there is a crucial difference between epiphenomenalism and physicalism for confirmation purposes when we assume that EUM has no bearing on confirmation. (In a way, we are really just applying the lessons from section III to the broad argument when we conclude this.) For every physical history of the world that remains a possible way things could turn out, there is only one determinate physicalist hypothesis that predicts it—the one that claims the actual qualitative event that would occur with each physical

neural base does occur with it. (Once again, this is because of physicalism's insistence on the metaphysical necessitation of qualitative event by physical neural base.) But there are numerous determinate epiphenomenalist hypotheses for each of these physical histories. Some of the epiphenomenalist hypotheses may have low intrinsic probabilities—e.g., the intrinsic probability of an epiphenomenalist hypothesis that predicted that the neural base actually associated with the sharp pain I feel when my arm is lacerated were associated instead with a complex visual experience. (Since there doesn't seem to be the right kind of information, and perhaps enough information, stored by the neural base to generate an experience of that kind without strange and unsimple happenings.) But as the discussion in connection with INPs and IPPs earlier in the work implied, there is reason to think many of the determinate epiphenomenalist hypotheses will have decent-sized intrinsic probabilities, and (more importantly) similar intrinsic probabilities. Moreover, most of these determinate epiphenomenalist hypotheses—even when we restrict our attention to ones that agree on laws, physical history of the world, and so forth—are guaranteed to conflict with whatever the evidence turns out to be, because they will make different predictions about the qualitative events instantiated. Thus, whatever the ultimate physiological evidence suggests about a physical history of the world, epiphenomenalism is sure to lose a greater portion of its overall volume than physicalism. Traditional defenders of broad evolutionary arguments were right about at least this much when they touted the standard conclusion.

But what about interactionism? To get a better feel for interactionism's prospects, we should at this point divide determinate interactionist hypotheses into two categories:

Lawful interactionist hypotheses (LIH)—interactionist hypotheses which claim that all qualitative events causing physical events do so in virtue of being governed by a specific kind of law—what we will call an "interaction law," which is a causal law from qualitative to physical.

Non-lawful interactionist hypotheses (NIH)—interactionist hypotheses which claim that not all qualitative events causing physical events do so in virtue of being governed by an interaction law.[204]

In addition to "downward" causation from qualitative event to physical event, all (currently held) interactionist views also posit "upward" causation from physical event to qualitative event, much as epiphenomenalist views do. There is no disagreement to speak of among interactionists on the question of whether these causal processes are lawful—all of them agree that upward causal processes are governed by laws. (We will call these laws "bridge laws.")

Let's consider LIH first. Hypotheses in this camp all posit upward bridge laws and downward interaction laws, and claim that these laws govern all instances where qualitative events cause happenings in the physical world. So, take a generic example of a situation where an external stimulus (a cut to the arm, say) causes (via causing various

[204] There are a number of philosophers enamored with a libertarian picture of free will that relies on agent causation. (E.g., O'Connor in O'Connor (2000).) These philosophers are typically dualists (often emergentists or even Cartesian substance dualists), but it is not obvious how to classify them in our schema. They seem to be interactionists and clearly believe in a kind of causation that is not law governed. At first blush, then, the most plausible position to place their views would be under the NIH heading, but they might be inclined to deny that qualitative events are relata in the non-lawful causal processes they embrace. I am not sure how to handle this problem and inventing a new category seems unnecessarily elaborate, so my inclination is simply to categorize their views as NIH, understanding that this categorization may not be perfect.

physical events in the nervous system) a characteristic qualitative event, which in turn causes a characteristic behavior (in this case, avoidance).[205] Pictorially, the situation would look like this:

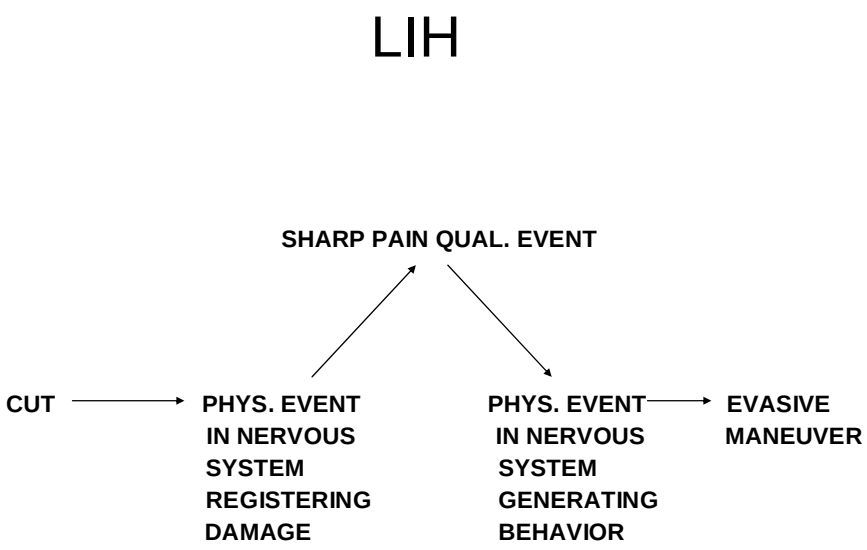

FIGURE 4

Now, consider the progression of physical events in this diagram. (Obviously, the diagram is grossly simplified and leaves out many intervening events, but it should be good enough to appreciate the point.) It either accords well with the Predictability Thesis or it doesn't.[206] It seems fair to conjecture that most LIH will not accord well with the

[205] Throughout the discussion, I assume that interaction laws preempt any physical laws that would otherwise hold. (In a later note, however, I briefly address the implausibility of interactionist hypotheses that countenance perfect overdetermination of physical to physical causal processes.) I ignore strange variant versions of interactionism that hold that physical and interaction laws compete with one another when the antecedents of each are satisfied.

[206] Obviously, much more than just the progression of physical events would need to be known to learn this with certainty—the outcomes in various counterfactual scenarios would also need to be known, for instance—but by 'accord' here, I just mean the level of fit with what we would expect if in fact the Predictability Thesis were true.

Predictability Thesis, and that most of the LIH intrinsic volume (and volume after background knowledge is taken into account) will be tied up in determinate hypotheses that claim something other than the Predictability Thesis. This is because bridge laws and interaction laws are not laws of physics—they are laws of some other fundamental scientific domain, perhaps psychology. As such, there is no reason to suppose that a determinate hypothesis that included them would be made complicated by having (in particular) interaction laws that led the physical world to transition in a way other than what we would expect if uniform physicalism were true. Non-uniform physicalist versions of physicalism, however, would suffer from added complication (relative to uniform versions of physicalism), because they would posit the existence of gerrymandered and *ad hoc* physical laws. But each determinate version of interactionism would simply have a set of physical laws, a set of bridge laws, and a set of interaction laws, and there would be no reason to suppose a set of laws which predicted qualitative events would violate the Predictability Thesis to be automatically less simple, even *ceteris paribus*, compared to a set of laws that predicted qualitative events would conform to the Predictability Thesis. And since there are many more ways for an interactionist hypothesis to violate the Predictability Thesis than to conform to it, it is natural to conclude that most of the general hypothesis's intrinsic volume would be invested in determinate hypotheses predicting something other than the thesis's fulfillment.

Consequently, any confirmation question involving the ultimate epistemic fate of physicalism or interactionism would likely turn on evidence about the truth of the

Predictability Thesis.[207] (Or, at the very least, this evidence would play an important role in answering such a question.) If we ultimately get overwhelming evidence in favor of the Predictability Thesis, then uniform physicalism will inherit a large portion of interactionism's intrinsic volume. And since, as I already argued, uniform physicalism is responsible for a large share of physicalism's overall volume, this will bode very well for physicalism generally. But, if we get overwhelming evidence against the Predictability Thesis, then the tables will be turned and interactionism will inherit much of physicalism's intrinsic volume.

Notice, however, that in this case, the news will not be as clearly and unmitigatedly good for interactionism as it would be for physicalism if the opposite sort of evidence were gathered (even if we assume that most of the intrinsic probability of interactionism is tied up in LIHs). This is because determinate versions of LIH will predict widely ranging physical transitions, whereas determinate versions of physicalism will tend to make predictions concentrated around certain Predictability Thesis-friendly histories. As a result, if the Predictability Thesis were found to be false, the specific way in which it were false would likely spell trouble for many mainstream LIHs as well, while if the Predictability Thesis were found to be true, comparatively few mainstream physicalist hypotheses would be in trouble. But nevertheless, evidence against the Predictability Thesis would still spell good news for LIH, and presumably for interactionism as a whole.

Before continuing on to a second problem for LIH and then to the NIH case, it is worth pointing out that although the Predictability Thesis is not often explicitly debated

[207] It will become important later that none of this discussion of the Predictability Thesis in connection with interactionism has presupposed anything about the correctness or incorrectness of objection (H).

in the philosophy of mind, it has been implicitly debated at great length. Typically, the implicit controversy manifest itself in debates over what is often called the "causal closure of the physical."[208] These debates center on empirical issues about the projectibility of the results of physical experimentation outside the brain to physical systems inside it, and these are the very issues that the Predictability Thesis takes a strong stand on.

While most philosophers seem to be of the opinion that the empirical findings strongly support the Predictability Thesis, the evidence we have is fairly meager and dissent is certainly possible.[209] (Though the most popular view of the evidence would have the consequence, if true, of strongly supporting physicalism.) As a result of philosophical and scientific disagreement about the evidence, as well as some of the issues raised above about deciphering the exact relevance of potential evidence we might obtain that undermined the Predictability Thesis, I will not attempt to say more here about what hypotheses are supported by the evidence we have in this sphere, though it is an interesting question in its own right just how and how much different potential Predictability Thesis evidence would impact the confirmation issue. I will rest content to have shown at least that evidence of this sort is far from irrelevant.[210] But even this

---

[208] Or sometimes the "causal exclusion of the physical," as we discussed before. (We had occasion to briefly examine causal closure/exclusion much earlier, back in Chapter 1.) Incidentally, the causal closure thesis is not equivalent to the Predictability Thesis—it neither entails that thesis, nor is it entailed by it. (It does not entail the Predictability Thesis because if a version of non-uniform physicalism were true, the causal closure thesis would be true while the Predictability Thesis would be false. And the Predictability Thesis doesn't entail it because if a version of interactionism were true that mimicked uniform physicalism in its predicted physical transitions, the Predictability Thesis would be true and the causal closure thesis false.) Nevertheless, the two theses are closely related.

[209] For a review of the evidence (a review favorable to a causal closure style theory), see Papineau (2002).

[210] I have heard many dualists who are not interactionists and physicalists complain informally about the *ad hocness* of interactionism, especially interactionist views that claim that qualitative events either overdetermine physical causal processes or preempt them in a way exactly in keeping with the Predictability Thesis. Although I cannot comment further here, it seems these philosophers may have more

modest finding is far subtler than the standard conclusion, which assumes that positing the causal efficacy of qualitative events is enough to secure confirmation in the light of the evidence. As we have seen, confirmation depends much more on the detailed predictions made about transitions in the physical world than it does on predictions about the mere existence of a causal relationship between qualitative events and physical ones.

As I alluded to above, there is another serious issue for interactionism (specifically, LIH) that the standard conclusion obscures, and this is an issue that parallels the one we saw for epiphenomenalism earlier in this section and in section III. (Its significance does also depend directly on the central assumption that we are making in this section—i.e., that objection (H) is misguided, and specifically that EUM is irrelevant for confirmation.) LIH posits the existence of both bridge laws and interaction laws, and as I argued previously in connection with epiphenomenalism (since epiphenomenalism also posits bridge laws), there is no reason to suppose that there is any more reason *a priori* (i.e., intrinsically) to suppose that the physical neural basis of (e.g.) a specific kind of sharp pain would produce that particular kind of pain rather than some other kind of qualitative event.[211] And the same goes, it seems, for interaction laws. Though a particular qualitative event causes a physical event φ in the actual world via some interaction law (let us suppose), there is no *a priori* reason why a law could not have been in place that would have led the same qualitative event to cause some subtly different physical event γ, which ultimately would have led to a very different behavior.

---

than just an ethereal intuition about the badness of *ad hoc* hypotheses—they may have a good empirical argument against these varieties of interactionism that can be expressed in confirmation-theoretic terms.

[211] Obviously there are limits to what qualitative events might be lawfully produced, though—if not limits of possibility, at least limits of the extent to which the intrinsic probability of such a law would be roughly equal to that of the one producing the sharp pain. (Recall the notes about the neural basis producing a complicated visual experience.)

Combining these two considerations, it seems that for each possible prediction about the transitioning of the physicalist world, there are numerous interactionist hypotheses that will entail it. These hypotheses will simply vary the bridge and interaction laws to make them work in harmony, producing the same physical transitions but mediating them with different qualitative events.

But if that is the case, return to our cut in the arm case once more. It could have instead looked like this:

LIH ALTERNATIVE

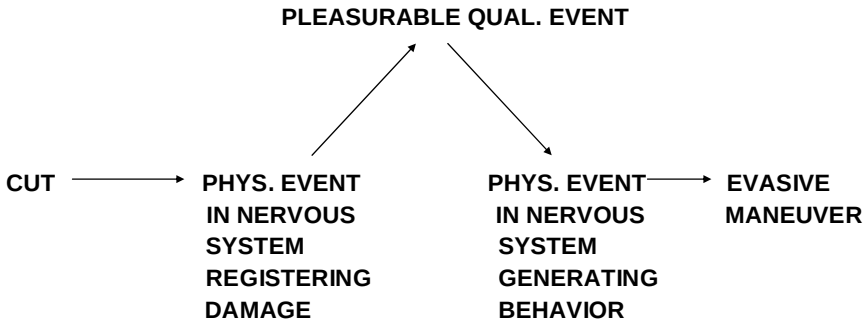

FIGURE 5

Thus, when we discover the actual phenomenological correlations and physiological transitions, we rule out this determinate version of LIH. But it is a perfectly respectable and mainstream version of LIH, and so a mainstream source of interactionism's overall intrinsic volume will be lost. Since there will likely be a large

number of similarly ill-fated LIHs, a large amount of interactionism's intrinsic volume will be lost. And since, as we discussed above, physicalism will have most of its intrinsic volume preserved (thanks to EUM's assumed irrelevance to confirmation), it will then inherit much of interactionism's volume, and epiphenomenalism's as well. Hence, in addition to any confirmation help physicalism gets in connection with evidence about the Predictability Thesis, it can count on help from this sphere as well. (And if it is hurt by evidence relevant to the Predictability Thesis, it can compensate for that in this sphere.) Notice that in addition to the obvious trouble this spells for interactionism and epiphenomenalism, it also spells trouble for the standard conclusion and its insistence that epiphenomenalism is the sole loser in broad evolutionary arguments.[212]

So, LIHs run into issues on two separate fronts—the first front has to do with the Predictability Thesis, and the second with its tolerance of determinate versions that allow for different qualitative events to play the same causal roles.[213] In conclusion, it is worth recapping a few things. First, issues about the Predictability Thesis may not be problems for LIHs, if the evidence (which is still quite meager) ultimately comes out a specific way—namely, that it fails to accord with what the Predictability Thesis would lead us to expect. Second, these issues surrounding the Predictability Thesis have nothing specifically to do with any temporary assumptions we are making about objection (H).

---

[212] Someone could, of course, defend the standard conclusion by claiming that we have good evidence against the Predictability Thesis (thus hurting physicalism and epiphenomenalism), but that the possibility of "mixing and matching" qualia on interactionism and epiphenomenalism hurts those two hypotheses and thus favors physicalism. And so, at the end of the day, only epiphenomenalism is uniformly hurt, while the other two general hypotheses trade punches. Nowhere in the literature is anything like this approach actually defended, though, nor is it plausible in the light of some of my later conclusions, so I won't address it further here.

[213] Of course, I don't mean to imply that the same determinate LIH will allow different qualitative events to play the same causal roles in the physical world, by having gerrymandered laws, where (e.g.) the same qualitative event plays a completely role in different people. (There may be such LIHs, but they will suffer from a low intrinsic probability.) I only mean to imply that different determinate hypotheses will make different predictions where this is concerned.

(They were only relevant to this section because they are general issues that face interactionism, and must be taken into consideration regardless of whether objection (H) is ultimately correct or not.) This will be important later. Third, the other variety of issue for LIHs is directly dependent for its significance on temporary assumptions we are making about objection (H)—specifically, that physicalism does not suffer from an analogous difficulty as the interactionist one being highlighted, because physicalism does not allow for the relevant kind of qualia "mixing and matching." And finally, given those assumptions, these issues are undoubtedly problems for LIHs, and interactionism generally.

Now that we hopefully have a good handle on the relationship between objection (H) and LIHs, let's focus on NIHs briefly. Many of the same issues that surfaced in connection with LIHs also surface with NIHs, only in a more extreme form. Because NIHs dispense with interaction laws (they posit only bridge laws), there will generally be little regularity in the causal patterns from qualitative to physical, and hence (since the Predictability Thesis strongly suggests regular, lawlike causal patterns in this domain) even more pronounced difference in the sorts of physiological transitions predicted. It is clear that the same observed physiological transitions that confirm physicalism will not also confirm NIH generally. And because NIHs posit the same sorts of bridge laws as LIHs and epiphenomenalist hypotheses, they will suffer from exactly the same sorts of worries where qualitative "mixing and matching" is concerned.[214]

[214] I won't venture a guess here as to how much of the overall intrinsic volume of interactionism is taken up by LIH and NIH respectively. This issue may be relevant ultimately in deciding how interactionism fairs once the evidence relevant to the Predictability Thesis comes in, but for our purposes, since LIHs and NIHs share many common general features germane to the context we are considering, I won't explore further. (There is little need because nothing about the overall "big picture" of our conclusion depends on it.)

This concludes the lengthy discussion of the prospects of broad evolutionary arguments conditional on objection (H) being misguided. It is clear that the reality where the confirmation of interactionism is concerned is much subtler than the standard conclusion would lead one to believe. The amount of confirmation or disconfirmation received by interactionism is heavily influenced by the particular kind of physiological transitions predicted (and the evidence gathered about those transitions), and the chances that the day of reckoning will be a happy one for both interactionism and physicalism is nil. Moreover, even if we set aside these issues, interactionism suffers from qualitative "mixing and matching" problems analogous to those faced by epiphenomenalism. Since these problems seem to be a central motivation for the broad evolutionary argument in the first place, it is curious that traditional proponents of that argument would fail to apply their alleged insight consistently, and extend it to interactionism as well.

A couple of very brief orders of business before moving on—first, very early in the work, in first formulating the narrower evolutionary argument, I asked readers to temporarily ignore any worries they might have that physicalism is *a priori* false. Many philosophers have certainly had these concerns, and they have given birth to a number of famous and influential arguments against physicalism—zombie arguments, knowledge arguments, and structural arguments. My reason for wanting the worries to be set aside was to allow us to examine physicalism and a variety of dualism (i.e., epiphenomenalism) empirically. But it is time now to begin stripping away the remaining artificial assumptions. For those absolutely convinced of the *a priori* falsity of physicalism, some natural questions might arise—can we use the evidence at our disposal to decide between interactionism and epiphenomenalism (i.e., the two remaining options), and if so, what is

the result? I will have a bit to say on this question in the morals and lessons section at the end of the chapter, but I would like to hold off on addressing it until then. This is primarily because, with the ultimate fate of objection (H) still up in the air, there is quite a bit more complication surrounding this question than there needs to be and ultimately (I hope) will be. Consequently, it makes sense to wait until after all the scaffolding has been removed to say anything more.

Second, we have still not addressed the lingering matter of NIC above. Recall that we could not simply dismiss NIC in the context of broad evolutionary arguments, because we were not entitled to assume NPLFA, only the weaker BPLFA. Hopefully, the intervening discussion of physicalism and the Predictability Thesis has put us in a position to more easily lay the matter to rest. The reason NIC is not an issue for broad evolutionary arguments is that, although there may be physicalist hypotheses that allow for special exceptions to the laws or for special laws governing physical events within the brain, these fall into two categories—either they are versions of uniform physicalism or not. Since *ex hypothesi* all such determinate versions of physicalism differ in their predictions from what is actually observed, if any are versions of uniform physicalism, then to the extent they are confirmed they will imply the falsity of the Predictability Thesis. This is because they will imply that significant counter evidence to it has been observed. (To appreciate why, it is important to notice that all of these determinate hypotheses will predict that highly unusual transitions have taken place, considering what they themselves represent the laws to be.) As a result, physicalism will be in trouble, but as we already saw, in this case physicalism will be in trouble anyway. The other category—versions that are not determinate uniform physicalist hypotheses—will not be

mainstream versions of physicalism (as we have already seen, since mainstream versions will tend to predict regular transitions), and so will have little impact on the overall intrinsic probability of physicalism. Thus, if they are ruled out, little of physicalism's overall probability will be lost and redistributed to other hypotheses.

V. The Truth of Objection (H)

We are now ready to directly examine what we have been taking for granted thus far—the falsity of objection (H) in its EUM form: i.e., the falsity of the claim that the lack of conceptual supervenience of qualitative on physical is without relevance where evolutionary arguments are concerned. (It may come as a relief to the reader that this stage of the process should be considerably less intricate than previous stages.) Recall that there is no issue about the epistemic separability of qualitative events and physical ones, even if physicalism is true. It is plainly apparent that even if physicalism is true, it is nonetheless conceivable in some sense that the physical neural base of an actual qualitative event be associated with some other qualitative event or no qualitative event at all. The only question is whether or not this has an impact on confirmation, and makes us judge the conditional probability of the evidence on physicalism (significantly) lower as a result. An affirmative answer to the question is essentially what objection (H) at its best is contending.

On the face of it, there doesn't seem to be any particularly good reason to doubt that EUM should have such an impact. After all, it seems that these alternate physicalist hypotheses are nevertheless epistemic possibilities, and ruling them out should have an adverse effect *ceteris paribus* on the likelihood of physicalism being true.

There is one notable objection to this proposal, though, which we will call the "Possibility Objection." Although it ultimately fails, it represents a potentially seductive challenge to the defender of objection (H) that is well worth exploring in depth.

It is easiest to appreciate the reasoning involved when we apply it to a concrete example. To keep things familiar, let 'C-fiber firing' denote whatever is the neural basis of the characteristic sharp pain associated with a knife wound to the arm. Also, we will restrict our attention to determinate physicalist hypotheses only.

The Possibility Objection acknowledges that it is epistemically possible that C-fiber firing be associated with something other than the sharp pain it is associated with.[215] But it then claims that the space of confirmation—to return to our metaphor, the circular region out of which the volumes arose—is not a space comprised of mere epistemic possibilities. Rather, it is a space comprised of those epistemic possibilities the agent believes to be metaphysical possibilities.[216] And unlike the determinate epiphenomenalist and interactionist options (which are dualist and so deny the metaphysical supervenience of qualitative on physical), the physicalist options are not all recognized as metaphysical possibilities from the get-go. In fact, it would be incoherent if they were, since this would imply the metaphysical supervenience of different

[215] One might get off the wagon at this point and deny, on *a priori* grounds, that there are any physicalist epistemic possibilities that involve any qualitative events whatsoever. But for the moment, I am not addressing individuals who believe physicalism can be shown to be false *a priori*, and I am assuming that there are epistemic possibilities corresponding to each way the world could turn out. I will have a bit more to say to *a priori* dualists in the lessons and morals section. For now, we are supposing we can make sense of the idea that physicalism and the presence of qualitative events are epistemically compatible.

[216] There is potentially the need to introduce a large amount of complexity into this basic picture if it is to do adequate justice to the intuitions being gestured at. (Degrees of confidence in the metaphysical possibility of entire maximal sets of worlds might need to be introduced, for instance, since it seems both that judgments about metaphysical possibility admit of degree, and also that strictly speaking, physicalist hypotheses are not compossible with dualist hypotheses.) However, since I will argue that even the broad outlines of the picture are misguided, I won't worry about addressing the need for this complexity.

qualitative events on the same physical supervenience base (e.g., both ecstatic pleasure and sharp pain on C-fiber firing).

And so, when the agent rules out various physicalist epistemic possibilities (e.g., the one that has ecstatic pleasure metaphysically supervening on C-fiber firing), this will not adversely affect the overall probability of physicalism, because these determinate hypotheses were not something that was contributing to that overall probability to begin with, since they were not in the overall space of confirmation to begin with.

To envision the way the Possibility Objection treats confirmation, return again to our giant circle. In the regions devoted to epiphenomenalism and interactionism, things proceed in basically the way discussed before. But in the region devoted to physicalism, there are no divisions into sub-regions for each determinate physicalist epistemic possibility, because these are not yet recognized as metaphysically possible. Rather, physicalism's region serves as a "place holder" for whatever determinate version of physicalism we ultimately settle on as compatible with the total evidence. Thus, physicalism's intrinsic probability is preserved as various determinate hypotheses are ruled out (and indeed, more volume is added as interactionism and epiphenomenalism have their determinate versions ruled out). All the while, the field of determinate physicalist hypotheses that are candidates for being actual dwindle, but without any adverse confirmation effects. Finally, at the (idealized) end of the day when all the evidence emerges, only one determinate version of physicalism is left standing, but this one version inherits all the intrinsic volume of physicalism[217] plus whatever volume came to the physicalist side after the renormalization from the lost volume along the way from

[217] Or, at the very least, all the volume physicalism had after the background knowledge was taken into account.

epiphenomenalism and interactionism. Undoubtedly, this results in tremendous overall confirmation for physicalism.

While interesting and perhaps seductive, the view of confirmation that undergirds the Possibility Objection is plagued by two major problems, the second even more serious than the first. The first problem is that if confirmation worked this way, then physicalism would only be getting confirmation via a cheap trick—whereas the other mind-body theories get disconfirmed when their individual determinate versions get ruled out by empirical evidence, physicalism doesn't, since its individual determinate versions don't get recognized as possible until they have been shown to be supported by the evidence! Thus, it might be suggested that physicalism starts with a much lower prior probability as a result, since it recognizes so many fewer possibilities to begin with. Consequently, the confirmation it receives only pulls it even with other options (at best).

The second and more serious problem is that confirmation just doesn't work this way—the space of confirmation is not a space of epistemic possibilities the agent believes are metaphysically possible. Rather, the space of confirmation is just as the Possibility Objection claims it isn't—it's a space of bare epistemic possibilities. The view that the space of confirmation is a space of epistemic possibilities the agent believes are metaphysically possible has immensely counterintuitive consequences. For example, presumably everyone believes the identity claim WATER=H_20 has been highly confirmed. And presumably the reason it has been highly confirmed is that it began with a certain intrinsic probability, and then as evidence was gathered and alternative identity claims were ruled out (such as, for example, WATER=XYZ), it inherited probability from these ruled out claims via renormalization. But if the Possibility Objection is

correct, then this can't be the correct diagnosis, since no coherently thinking agent would recognize the metaphysical possibility of all the competing identity statements at once, since each is a metaphysically necessary truth if a truth at all, and the truth of each one is incompatible with the truth of the others. The allegedly correct diagnosis is rather that a more general claim, something like WATER IS IDENTICAL TO A PHYSICAL SUBSTANCE, was confirmed because its intrinsic probability was maintained as the evidence was taken into account while the intrinsic probability of other options (WATER IS AN OPTICAL ILLUSION (e.g.); WATER IS A CHEMICAL MIXTURE) was siphoned off. All the while, potential determinate versions of WATER IS IDENTICAL TO A PHYSICAL SUBSTANCE were being narrowed down, till only the one remained.

Convoluted as this account is, it gets even worse when we contemplate the confirmation of the specific proposition WATER=H_20. Although the convoluted account at least produces the right answer to the question "was the proposition WATER IS IDENTICAL TO A PHYSICAL SUBSTANCE confirmed?" (i.e., yes!), it cannot produce the right answer to the question of whether WATER=H_20 was confirmed. Rather than giving the obviously correct answer that everyone agrees on—i.e., that the proposition was confirmed—it must claim that WATER=H_20 had no intrinsic probability, and only can be said to have a probability at all (i.e., to use the metaphor, a share of the volume) when it is the only determinate option left standing among the versions of WATER IS IDENTICAL TO A PHYSICAL SUBSTANCE. (Recall that confirmation is essentially a raising of the probability of a hypothesis by considering the evidence. But if WATER= H_20 had no probability along the way, then there was no probability to raise.)

The ridiculousness of this conclusion is too much to stomach. The Possibility Objection is false, and since there are no other challenges to EUM's relevance to confirmation, we must conclude that objection (H) is in fact correct. The issue we must turn to now is the relevance of this for evolutionary arguments.

VI. The Upshot—Evolutionary Arguments Fail

As I've alluded to throughout the chapter, this conclusion about objection (H) is of crucial importance. It spells doom for both narrow and broad evolutionary arguments. Although the inferences may already be clear, it is worth spelling them out explicitly. I will do this first for the narrower argument, because the reasoning is easier both to formulate and appreciate. Once we have dealt with this easier case, we will move on to the broad argument.

Back in section III, I defended the conditional conclusion for the narrower evolutionary argument that if objection (H) was successful and EUM had a bearing on confirmation, then physicalism does no better job leading us to expect the evidence we actually find than epiphenomenalism. Now we have established the antecedent of that conditional—objection (H) is successful. Hence, physicalism is not confirmed over against epiphenomenalism.

This is especially easy to see in the narrow case because we are able to take for granted the history of the physical world (extending NPLFA slightly). Thus, all the hypotheses in play agree on the transitioning of physical entities—the only differences involve the posited correlations of qualitative events with physical events. But for every determinate epiphenomenalist epistemic possibility, there will be a corresponding

physicalist epistemic possibility. For example, just as there is an epiphenomenalist possibility that connects C-fiber firing with a certain kind of pleasure, there will be a physicalist possibility that does the same. When the actual observations are made and determinate versions of epiphenomenalism and physicalism are ruled out, there will be a symmetric process of ruling out on each side. And since there is no principled reason to suppose that the determinate epiphenomenalist hypotheses that are ruled out will be responsible for a greater share of epiphenomenalism's overall prior probability than the determinate physicalist hypotheses will be physicalism's overall prior probability (and vice-versa), there is no reason to suppose either general hypothesis will be confirmed or disconfirmed relative to the other.

Now when it comes to broad evolutionary arguments, the relevance of objection (H) is slightly harder to see, but only slightly. Essentially, what the truth of objection (H) does is strip physicalism of its ability to take advantage of the metaphysical contingency of the correlation between qualitative events and physical ones on epiphenomenalism and interactionism. The metaphysical contingency of these correlations on these views, and the metaphysical necessity of them on physicalism, is of no significance to the argument. Because the correlations are epistemically contingent on all the views, and because the space of confirmation is the space of epistemic possibility (as we saw above), any time observation rules out an epistemically possible correlation, there will be analogous loss of volume by all the general hypotheses. There may be room for subtle differences between the hypotheses (in particular, between interactionism and the other options, owing to interactionism's added laws), but if they exist, these differences will be very subtle

indeed, hardly enough to confidently ground any sort of argument against any of the views.

The only considerations that have a hope of aiding us in the process of deciding between the various general views are (as we have seen earlier) evidence relevant to the Predictability Thesis, and *a priori* considerations,[218] since surely if physicalism is ruled out *a priori* (or alternatively, is made palatable *a priori*), this will have significant effects on the intrinsic probabilities of the determinate physicalist options, and also on the intrinsic probabilities of the determinate dualistic hypotheses, since probability in this setting is a zero sum game (as the volume metaphor should make clear).[219]

VII. Morals and Lessons

We have now come to the end of our journey and, alas, have only a negative conclusion to show for it. But we have gained insights along the way, and so we needn't bemoan our lack of a positive conclusion too greatly. What I would like to do in this concluding section is sum up some of the main general lessons we have learned from our discussion—not just the now familiar conclusions, but also the implications of those conclusions, some of which have been discussed prominently and some of which have not.

---

[218] I am counting as *a priori* here more than just inferential relations between concepts and the like. I am also including arguments and intuitions about the limitations of (e.g.) conceivability as a guide to possibility.

[219] There is another type of evidence that could potentially play a role. If it were found that qualitative events did not correlate smoothly with any neural basis, this would be evidence for dualism over physicalism. This is because only dualism allows for the possibility of this sort of variation, though only intrinsically far-fetched versions of dualism predict this. In any case, virtually every indication we have suggests that we will not find this, and almost no one (physicalist or dualist) suggests otherwise, so I will not bother to consider this possibility further.

To begin with the major findings, we concluded that both narrow and broad evolutionary arguments ultimately fail, because objection (H) succeeds—physicalism does not lead us to expect the evidence any better than epiphenomenalism. In fact, it is not clear at present that any of the general theories do a better job than any of the others at predicting the evidence we actually find! In keeping with this theme, we also found that the standard conclusion was false—there is no reason to suppose, even if objection (H) is false, that physicalism and interactionism are both confirmed, and epiphenomenalism alone disconfirmed.

Another important lesson was that the intrinsic positivity and negativity of the qualitative events that were correlated with the various neural events did not drive the argument toward its conclusion. What really mattered all along was rather that some qualitative event was stably correlated with some physical neural event, but which one was of little significance—from there, it was the metaphysical contingency of the supervenience relation according to epiphenomenalism and interactionism, and the metaphysical necessity of it according to physicalism, that allowed physicalism to enjoy the crucial dialectical advantage where qualitative mixing and matching was concerned. (All of this was conditional, of course, on the falsity of objection (H).) Granted, this result only became clear when we moved from considering broad evidence (classifying phenomenology as only INP/IPP) to considering the determinate evidence—specific phenomenology and physiological transitions. But this shift was highly principled, since it is standard practice in abductive arguments to formulate evidence as specifically as possible, and doing otherwise can lead to very counterintuitive results. Moreover, in the light of our more precise analysis of the confirmation issues surrounding the general

argument strategy in this chapter, it is difficult to see how an argument using less fine-grained evidence would have faired any better. Although by using less fine-grained evidence fewer determinate possibilities will be ruled out, it is hard to convince oneself that this would result in a significantly different distribution of lost probability (between the three general theories) than the one we obtain by taking into account the more fine-grained evidence. Thus, there is no reason to suppose taking account of only the broader evidence would have helped the prospects of the general argument strategy.

Similarly, at the end of the day, evolutionary evidence (and even evidence about the continued survival of presently living organisms) does not play a crucial role in driving the argument (either the broad argument or the narrower argument) toward its conclusion. This evidence is essentially just background knowledge, and even if it isn't background knowledge, there doesn't seem to be any reason to believe that one of the general theories has an overall advantage (or overall disadvantage) in being able to account for it. None of the general theories had any unique features that systematically gave them an advantage in leading us to expect this evidence, and consequently none gain appreciably from it.

We did learn, however, that certain types of evidence could play a significant empirical role in deciding between different mind-body theories. Evidence relevant to the truth of the Predictability Thesis, although currently meager, could hold the key to any empirical resolution of disputes about the mind-body problem. The reason is that, as we have seen, physicalism and epiphenomenalism as general theories lead us to expect something very different from what interactionism as a general theory leads us to expect. Hence, the situation is ripe for experimentation to settle the matter.

To return to an issue left hanging above, this method could be especially useful for those convinced that physicalism can be shown to be false *a priori*. Although evolutionary evidence and evidence about correlations between qualitative events and physical ones won't help someone in this position, evidence about the Predictability Thesis could come in very handy in the effort to discern just what type of dualism is true.

Empirical arguments can only take us so far, though. If the Predictability Thesis is ultimately well confirmed, then considerations like the ones we've been examining throughout this work won't help us to decide between epiphenomenalism and physicalism. Because both general hypotheses predict the same basic physical transitions (and don't differ in their predictions about things like qualitative event-physical neural event correlations), no empirical observations will resolve any disputes that arise between proponents of the competing theories. At that point, there will be no hope but to return to our armchairs and go back to philosophizing the old fashioned way.

Appendix—Glossary of Commonly Used Acronyms

This appendix is designed to help the reader to more conveniently keep track of acronyms commonly used throughout the dissertation. The acronyms are listed in alphabetical order, and each includes an explanation of what the acronym's individual letters stand for, the chapter in which the acronym is introduced, and an informal definition.

BPLFA—Broad Physical Law Fixing Assumption—Chapter 4—A claim made in connection with the broad evolutionary argument (i.e., the argument that considers epiphenomenalism, interactionism, and physicalism all as dialectical options). It states that the argument will treat as background knowledge the initial physical conditions of the world, the physical laws governing physical systems that do not include brains, and the progression of the physical world up to the time that sentient beings first appeared.

EUM—Epistemically Uncausally Mediated Version (of the unbreakability objection)—Chapter 4—The claim that qualitative events do not follow with epistemic (i.e., conceptual) necessity from the arrangement of physical events if physicalism is true.

INP—Intrinsically Negative Phenomenology—Chapter 1—Phenomenology which feels "bad" or "nasty." The badness is not reducible to facts about desires or aversions.

IPP—Intrinsically Positive Phenomenology—Chapter 1—Phenomenology which feels "good" or "pleasant." As with INP, this phenomenology is not reducible to facts about desires.

LIH—Lawful Interactionist Hypotheses—Chapter 4—Interactionist hypotheses which claim that all qualitative events causing physical events do so in virtue of being governed by causal laws from qualitative to physical.

MUM—Metaphysically Uncausally Mediated Version (of the unbreakability objection)—Chapter 4—The claim that qualitative events do not follow with metaphysical necessity from the arrangement of physical events, if physicalism is true. (This is an obviously false claim.)

NAC—Nomically Accessible Causally Mediated Version (of the unbreakability objection)—Chapter 4—The claim that indeterminism in the actual physical laws could have produced human organisms with very different physiological constructions from the actual ones, and very different phenomenology/behavior correlations, if physicalism is true.

NIC—Nomically Inaccessible Causally Mediated Version (of the unbreakability objection)—Chapter 4—The claim that different physical laws could have produced human organisms with very different physiological constructions from

the actual ones, and very different phenomenology/behavior correlations, if physicalism is true.

NIH—Non-lawful Interactionist Hypotheses—Chapter 4—Interactionist hypotheses which claim that not all qualitative events causing physical events do so in virtue of being governed by a law from qualitative to physical.

NIIC—Nothing Interesting Intrinsic in Common—Chapter 2—The thesis that IPPs have no fairly joint-carving intrinsic properties in common in all situations and worlds where they are IPPs. (One could formulate an analogous thesis for INPs.)

NPLFA—Narrow Physical Law Fixing Assumption—Chapter 4—An assumption used in connection with narrower evolutionary arguments (i.e., ones that only consider epiphenomenalism and physicalism). It states that the initial physical conditions of our universe and all the laws of nature that actually govern the progression of physical events and the transitions of physical systems can be treated as metaphysically and epistemically necessary truths for the purposes of argument.

BIBLIOGRAPHY

Alexander, S., 1920. *Space, Time, and Deity*. London: Macmillan.

Armstrong, D.M., 1989. *Universals: An Opinionated Introduction*. Boulder, CO: Westview Press.

Aydede, M., 2000. An Analysis of Pleasure Vis-à-Vis Pain. *Philosophy and Phenomenological Research* 61: 537-570.

Bayne, T., 2001. Chalmers on the Justification of Phenomenal Judgments. *Philosophy and Phenomenological Research* 62: 407-419.

Bealer, G., 2002. Modal Epistemology and Rationalism. In T. Szabo Gendler and J. Hawthorne, eds. *Conceivability and Possibility*. Oxford: Oxford University Press.

BonJour, L., 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

BonJour, L., 1999. The Dialectic of Foundationalism and Coherentism. In J. Greco and E. Sosa, eds. *The Blackwell Guide to Epistemology*. Malden, MA: Blackwell.

Broad, C.D., 1925. *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.

Buck, R., 1976. *Human Motivation and Emotion*. Chichester, UK: John Wiley & Sons.

Chalmers, D., 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D., 2003. The Content and Epistemology of Phenomenal Belief. In Q. Smith and A. Jokic, eds. *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press.

Chalmers, D., 2004. The Representational Character of Experience. In B. Leiter, ed. *The Future for Philosophy*. Oxford: Oxford University Press.

Chalmers, D., 2006. Two-Dimensional Semantics. In E. Lepore and B. Smith, eds. *Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press.

Conee, E. and Feldman, R., 1998. The Generality Problem for Reliabilism. *Philosophical Studies* 89: 1-29.

Davidson, D, 1970. Mental Events. In L. Foster and J. Swanson, eds. *Experience and Theory*. London: Duckworth.

Dennett, D.C., 1978. Why You Can't Make a Computer that Feels Pain. In his *Brainstorms*. Cambridge, MA: MIT Press.

Draper, P., 1989. Pain and Pleasure: An Evidential Problem for Theists. *Nous* 23: 331-350.

Draper, P., 1997. Evolution and the Problem of Evil. In L. Pojman, ed. *Philosophy of Religion: An Anthology, 3rd edition.* Belmont, CA: Wadsworth.

Eccles, J. and Popper, K., 1977. *The Self and its Brain: An Argument for Interactionism*. Berlin: Springer-Verlag.

Fodor, J., 1990. Making Mind Matter More. Reprinted in *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fumerton, R., 1995. *Metaepistemology and Skepticism*. Lanham: Rowman and Littlefield.

Gertler, B., 2001. Introspecting Phenomenal States. *Philosophy and Phenomenological Research* 63: 305-328.

Goldman, A. I., 1976. What Is Justified Belief?. In G.S. Pappas, ed. *Justification and Knowledge*. Dordrecht: Reidel.

Hawthorne, J, 2001. Causal Structuralism. *Nous* 35: 361-378.

Hawthorne, J., 2004. Why Humeans are Out of Their Minds. *Nous* 38: 351-358.

Hawthorne, J., forthcoming. Direct Reference and Dancing Qualia.

Horgan, T., and Tienson, J., 2002. The Intentionality of Phenomenology and the Phenomenology of Intentionality. In D. Chalmers, ed. *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press.

Howson, C. and Urbach, P., 1996. *Scientific Reasoning: The Bayesian Approach, 2nd edition.* Chicago: Open Court.

Huxley, T, 1874. On the Hypothesis that Animals are Automata. Reprinted in *Collected Essays*. London, 1893-94.

Jackson, F., 1972. Review of K. Campbell, *Body and Mind*. *Australasian Journal of Philosophy* 50:77-80

Jackson, F., 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-136.

Jackson, F., 1986. What Mary Didn't Know. *The Journal of Philosophy* 83:291-95.

Jackson, F., 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

James, W., 1879. Are We Automata?. *Mind* 4:1-22.

James, W., 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Joyce, J., 2005. How Probabilities Reflect Evidence. *Philosophical Perspectives* 19: 153-178.

Kim, J., 1998. *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kripke, S., 1972. Naming and Necessity. In G. Harman and D. Davidson, eds., *The Semantics of Natural Language*. Dordrecht: Reidel.

Kripke, S., 1982. *Wittgenstein on Rule-Following and Private Language*. Cambridge, MA: Harvard University Press.

Leslie, J., 1989. *Universes*. London: Routledge.

Lewis, D., 1983. New Work for a Theory of Universals. *The Australasian Journal of Philosophy* 61: 343-77.

Lewis, D., 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lindahl, B. I. B., 1996. Consciousness and Biological Evolution. *The Journal of Theoretical Biology* 187: 613-629.

Loewer, B., 2001. Review of *Mind in a Physical World*. *The Journal of Philosophy* 98: 315-321.

McLaughlin, B., 1995. Varieties of Supervenience. In E. Savellos and U. Yalcin, eds. *Supervenience: New Essays*. Cambridge: Cambridge University Press.

McLaughlin, B., 2006. Mental Causation. In D. Borchert, ed. *Encyclopedia of Philosophy, 2nd edition*. Detroit: Macmillan

Meacham, C., unpublished. Sleeping Beauty and the Dynamics of De Se Belief.

Melzack, R; Wall, P.; et al., 1994. Textbook of Pain, 3rd edition. Edinburgh, NY: Churchill Livingstone.

Nozick, R., 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press

O'Connor, T., 2000. *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.

Papineau, D., 2002. *Thinking About Consciousness*. Oxford: Clarendon Press.

Plantinga, A., 1993. *Warrant: The Current Debate*. Oxford: Oxford University Press.

Price, D., 1999. *The Psychological Mechanisms of Pain and Analgesia*. Seattle: IASP Press.

Robinson, W. S., 1982. Causation, Sensations, and Knowledge. *Mind* 91: 524-540.

Robinson, W.S., 2003. Epiphenomenalism. In *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/epiphenomenalism/#Natural.

Robinson, W. S., 2004. *Understanding Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Robinson, W. S., forthcoming a. "What is it Like to Like?"

Robinson, W. S., forthcoming b. "Evolution and Epiphenomenalism."

Romanes, G. J., 1896. *Mind and Motion, and Monism*. London: Longmans, Green, and Co.

Rosenberg, G., 2004. *The Nature of Mind*. Oxford: Oxford University Press.

Russell, B., 1910. Knowledge By Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society* 11: 108-128.

Russell, B., 1927. *The Analysis of Matter*. London: Kegan Paul.

Schroeder, T., 2001. Pleasure, Displeasure, and Representation. *Canadian Journal of Philosophy*, 31: 507-530.

Shoemaker, S., 1980. Causality and Properties. In Peter Van Inwagen, ed. *Time and Cause*. Dordrecht: Reidel.

Sider, T.R., 1999. Global Supervenience and Identity Across Times and Worlds. *Philosophy and Phenomenological Research* 59: 913-937.

Sober, E., 2005. The Design Argument. In *The Blackwell Guide to the Philosophy of Religion*. Oxford: Blackwell.

Spencer, H., 1870. *First Principles of a New System of Philosophy, 2nd ed*. New York.

Spencer, H., 1871. *Principles of Psychology*. New York.

Spencer, H., 1883. *Principles of Biology.* New York: D. Appleton and Company.

Stalnaker, R., 1996. Varieties of Supervenience. *Philosophical Perspectives* 10: 221-241.

Stich, S., 1983. *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Swinburne, R., 1973. *An Introduction to Confirmation Theory*. London: Harper and Row.

Taylor, R, 1992. *Metaphysics: 4th Edition*. Englewood Cliffs, NJ: Prentice Hall.

Tooley, M., 1987. *Causation: A Realist Approach*. Oxford: Oxford University Press.

Trigg, R, 1970. *Pain and Emotion*. Oxford: Clarendon Press.

White, R., 2000. Fine-Tuning and Multiple Universes. *Nous* 34: 260-276.

Williamson, T., 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Yablo, S., 1992. Mental Causation. *The Philosophical Review* 101: 245-280.

Curriculum Vita

Joseph Anthony Corabi

Universities Attended

September 1997—May 2001	Saint Joseph's University, B.A. in Philosophy
September 2001—October 2007	Rutgers University, Ph.D. in Philosophy

Positions Held

September 2001—May 2004	Excellence Fellowship, Rutgers University
September 2004—May 2007	Teaching Assistantship, Rutgers University
August 2006—May 2007	Visiting Asst. Professorship, Saint Joseph's Univ.