

Physicalism and Evidence

by

Kyle Motsinger

Department of Philosophy
Duke University

Date: _____

Approved:

Karen Neander, Supervisor

Owen Flanagan

Alex Rosenberg

Thesis submitted in partial fulfillment of
the requirements for the degree of Master of Arts in the Department of
Philosophy in the Graduate School
of Duke University

2012

ABSTRACT

Physicalism and Evidence

by

Kyle Motsinger

Department of Philosophy
Duke University

Date: _____

Approved:

Karen Neander, Supervisor

Owen Flanagan

Alex Rosenberg

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Arts in the Department of
Philosophy in the Graduate School
of Duke University

2012

Copyright by
Kyle Motsinger
2012

Abstract

One popular characterization of physicalism is as an empirical prediction about the theoretical objects and properties of future science. It is argued that physicalism so characterized cannot be justified unless an individual has idiosyncratic standards of inductive evidence. An alternative to this characterization of physicalism is a project that justifies physicalism according to the current science: all theoretical objects and properties can either be reduced to or eliminated in favor of the objects and properties of the current physical sciences. The extent to which this type of physicalism can be justified depends on an individual's epistemic rules and prioritization thereof.

Contents

Abstract.....	iv
1. Epistemic Rules and their Diversity.....	1
2. Two Questions of Physicalism.....	14
3. Physicalism as an Empirical Prediction.....	20
4. Bickle's Physicalism as a Non-inductive Project.....	34
Bibliography.....	44

1. Epistemic Rules and their Diversity

Our epistemic lives can be broadly characterized and prioritized in various ways. Descartes, of course, did a thought experiment to try to ground everything we believe, while Humean skepticism dismayed of having any actual knowledge but was content with feelings of knowledge. A clear concern with skepticism reigned in the modern era, but pragmatism shifted the focus partially away from skepticism onto a different question: what does our overall belief system look like, and what should it look like? This is well articulated in James's "The Will to Believe". We have to decide, James says, where we want to fall along a particular epistemic continuum. We can either try to believe as many true things as possible, try to avoid believing as many false things as possible, or fall somewhere between the two extreme. In James's day, the extreme of avoiding error was represented by his contemporary "enfant terrible" (as James called him), W.K. Clifford, and his famous dictum, "it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence."¹ James took himself to be somewhere near the other end of the spectrum, stating that "a rule of thinking which would absolutely prevent me from acknowledging certain kinds of truth if those kinds of truth were really there, would be an irrational rule."² James contended that this permits believing propositions for which we do not have evidence (but, importantly, which we do not have evidence against, either), including particular items of religious faith, insofar as we are psychologically capable of believing the particular propositions.

¹ Clifford 1877, p. 295

² James 1911, p. 28

This perspective is novel because it implies that propositions aren't justified *tout court*, but rather that whether we are justified in believing a proposition may depend on individual epistemic preferences: a proposition that is justified for James may be completely unjustified for Clifford. Now this isn't what James himself believes, as he sees Clifford and others on the error-avoidance end of the epistemic spectrum as irrational; however, for our purposes, the takeaway from James is this: epistemic justification does not float freely from the individual, but depends on the epistemic preferences of the individual.

Jump ahead fifty-five years, and we arrive at a new twist on James's philosophical innovation exemplified in Quine's "Two Dogmas of Empiricism". At the beginning of the final section of the paper, Quine introduces a powerful metaphor: "...the totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges."³ In a phrase Quine would coin later, each person's epistemic life can be characterized as a web of belief; the edge of the web is an individual's myriad experiences, and the inner portion the complex relations between various beliefs. What guides the development of the web is, of course, each individual's experiences, but also various preferences that dictate how one is to adjust the web (i.e., which beliefs to keep, drop, or revise) in response to experience. So far, we simply have a picturesque metaphor that James might have been perfectly comfortable with. What is new (and is

³ Quine 1953, p. 42

actually an inheritance from earlier logical positivists and not original with Quine) is that the rules for revising the belief go beyond the simple continuum of believing truth and avoiding error.

More specifically, there are no strict constraints on what belief-revision rules may be used. This was more or less explicitly stated by Carnap in a similar (but not identical) context: "To decree dogmatic prohibitions of certain linguistic forms instead of testing them by their success or failure in practical use, is worse than futile; it is positively harmful because it may obstruct scientific progress."⁴ Here, Carnap is arguing against restricting the use of linguistic frameworks for various inquiries, but without a viable analytic-synthetic distinction (as Quine argues in "Two Dogmas"), this becomes equivalent to arguing against restrictions on rules for belief revision.

Clearly, however, there will be many belief-revision rules that no person would ever use in practice. A rule to only believe what one reads in the Weekly World News (and nothing else) would not produce many true beliefs. Even a rule to believe on what one reads in the New York Times (and nothing else) would leave an individual severely lacking, as there are countless everyday beliefs that we need to survive that aren't in the pages of the Old Gray Lady (that your car is out of gas, for instance). Of course, these rules in question concern the source of beliefs, and all people at least use sense perception (broadly construed) as a source of belief. This is common ground that we can take for granted, but the question of how to integrate beliefs from sense perception still remains.

⁴ Carnap 1956, p. 221

Not everything we see, hear, smell, or, particularly, read will fit nicely with our pre-existing beliefs, even if all of those pre-existing beliefs originated in perceptual experience. The most obvious cases of conflict are when two beliefs present contradictory information upon which we must act: the conference website said that it starts at 10:00 am, but a recent email said that it now starts at 10:30 am. The old belief and the new information conflict; in the case of the conference's starting time, most people will obviously update their belief on the rule more recent information from a source trumps older information from that same source (*ceteris paribus*, of course). Likewise, no one is going to use the rule that states to keep the belief that can be most succinctly stated in Latin. Nothing is wrong with such a belief-revision rule *per se*, but most people do value true beliefs (particularly true beliefs that pertain to punctuality), and such a rule does not reliably produce true beliefs.

So what does this have to do with physicalism (or even ontological questions in general)? The debate over physicalism, as with philosophical debates more broadly, is plagued by unstated epistemic commitments; a preference for one belief-revision rule over another can lead to diverse conclusions given the same set of basic facts (illustrating this will be a task undertake below). Arguing against a particular opponent might be fruitless if you share a different set of epistemic commitments than your opponent, and if this is the source of the disagreement, then either you must shift the debate to the appropriate epistemic commitments (which is a practical argument that must ultimately assert that your opponent's epistemic commitments are not the best commitments to fulfill their preferences) or simply agree to disagree. Because epistemic

commitments (much less preferences) are rarely explicitly stated before individuals start elaborating an argument, one may have to do some detective work (or at least assumptions about their commitments) to figure out if the argument is worth considering. This isn't to say that the argument might not be interesting for other reasons, but that ultimately it will fail to be persuasive because you share different epistemic commitments than those who do find it persuasive.

For a toy example, we can refer back to the conflict between James and Clifford. For James, belief in the existence of God wasn't so much the product of a formal argument, but had he been asked to explain why he believed in God, he might have said something along the following lines: I have had empathetic and emotional experiences in a certain religious context, and those experiences produced in me a belief that there is something that is like the Christian God. Such a justification holds zero persuasive force for Clifford. Experiences of overwhelming emotion do not qualify as evidence for belief, and given his earlier dictum ("it is wrong always, everywhere..."), such an emotional experience would not be sufficient ground for belief (at least to justify a belief; psychologically, he may acquire the belief in spite of himself). Clearly James does not use the same epistemic rules Clifford; thus, there is no ground on which they can discuss this particular reason for believing in God. Rather, they must engage in a discussion about which epistemic rules are appropriate. Ultimately, given that, ideally, particular epistemic rules are used because of the specific preferences an individual has, a discussion about the proper epistemic rules may end in an irresolvable disagreement.

All is not lost, however, as there will still be some basic epistemic rules shared across individuals (revision via perception being a major one), and there will still be some beliefs that are accepted by multiple individuals even across a diversity of epistemic rules. One such rule that every discussant will share is a rule to reject contradictions. Every philosopher learns this rule by heart in their first logic class: from a contradiction anything can be derived. Thus, contradictions are bad from an inferential standpoint, but they are also bad from a conceptual standpoint because it's not clear how a proposition could be both true and false at the same time. The kind of revision to one's conceptual scheme to accommodate the possibility that a proposition could be both true and false would be enormous and might not leave much recognizable in its wake. Unfortunately, even an epistemic rule as simple as "reject all contradictions" is too simple when universally applied to one's beliefs. Rejecting one contradiction may lead to other contradictions in the revision, and depending on one's current set of beliefs, there may be no way to eliminate all contradictions and still minimally satisfy other epistemic rules. One could (theoretically, not psychologically) believe nothing and thus avoid all contradictions, but all individuals have clear preferences that dictate having some beliefs. Thus, it may not be possible for a person to hold to a very strong rule about contradictions, such as "eliminate any and all contradictions no matter the consequences of to the total set of beliefs." A person may, in a sense, be stuck at a "local maximum" where the best one can do given the other epistemic rules one holds is to have a few contradictions somewhere in the complete set of beliefs. A weaker, and doubtless more widely used, rule would be to minimize the number of contradictions

insofar as one does not significantly violate other epistemic rules. This might even entail consciously realizing that you hold a contradictory belief, but not revising it because to do so would do great violence to your overall set of beliefs. It's doubtful that we're going to identify too much variance in epistemic rules over the minimization of contradictions, but it should be recognized as a possibility.

Another broad rule that everyone will share to some extent concerns prediction: beliefs that result in successful predictions are better than beliefs that result in unsuccessful predictions. Because all people are individual organisms that have desires and must interact with the world to fulfill those desires, everyone will employ this epistemic rule to some extent. Most of our beliefs that are formed in accordance with this rule are those near the "edges" of our web of belief: those that connect quite directly with experience. The great majority of the beliefs that are used in prediction are minimally theoretical everyday beliefs: that there is milk in the fridge, or that your watch is five minutes fast. Clearly these beliefs, insofar as they are formed *using* epistemic rules, are not formed only using a rule to maximize the predictive success of our beliefs. In fact, such a rule may be more regulative of other rules and belief-acquisition methods than it is of directing the acceptance and rejection of beliefs themselves. Such a rule may direct us to accept basic enumerative induction as a further epistemic rule to use in many situations. Thus, we can not only encounter various prioritizations of independent epistemic rules, but also various hierarchies of epistemic rules. Most everyone will use simple induction for some things, but not everyone will require randomized controlled trials as a necessary rule for forming beliefs on, say, the

efficacy of medicines. The epistemic rules that guide the revision of beliefs used in making predictions can thus vary across individuals. Some individuals will be more stringent in their epistemic standards, whereas others will be more lax, and this variance can result from a huge variety of preferences (not the least of which is the aforementioned contrast between James and Clifford: maximizing true beliefs vs. minimizing false beliefs).

As philosophers are wont to point out, however, prediction and explanation are not the same thing. I can predict that the sun will rise in the morning because it has risen every morning I have been alive, but that doesn't entail that I can explain why the sun will rise in the morning. Given this, explanation can still be seen as closely related to prediction. Often, being able to explain why something occurs comes with being able to predict when/how something will occur. Clearly this isn't always the case. I can explain why heavier-than-air flight is possible, but I can't predict how any particular airfoil will perform. I can also explain why the sun rises in the morning, but that explanation is not essential to my predictive ability. With a sufficiently formalized theory, however, I can use it to predict events. Explanation is thus not coextensive with prediction, nor even a subset of prediction. As a result, the epistemic rules that an individual uses govern prediction- and explanation-related belief revisions may be different. A clear example is again the case of the sun rising: the rule for forming predictions about the sun rising is completely different than the rule for explaining that the sun will rise. The former is a

case of simple enumerative induction; the latter is a case of something we can label (while doing little justice to the complexities involved) “the scientific method”.

Unlike prediction, explanation seems to entail that a person has a kind of theory (formal or informal) about what is being explained, and explanations that produce predictions generally involve a theory containing objects with causal powers. This has direct consequences for ontological beliefs. Epistemic rules that (directly or indirectly) govern the revision of explanation-related beliefs also govern the revision of some ontological beliefs. Not all ontological beliefs are formed this way; clearly some people have ontological beliefs that have not been formed using explanation-related epistemic rules at all (e.g., some beliefs about the afterlife), but there are also countless ontological beliefs for which it would be quite a stretch to say that one has used anything resembling the scientific method to form them (think of the belief there exists a paper that you are currently reading). Yet for those ontological beliefs that are formed or revised via explanation-related epistemic rules, a diversity in such rules could result in disagreement on ontologies and a subsequent intractable debate over those ontologies.

So far, we have articulated four types of epistemic rules: maximizing true beliefs vs. minimizing false beliefs, minimizing the number of contradictions within an individual’s set of beliefs, rules for ensuring successful prediction, and rules governing explanation. The last type of epistemic rules that we will cover (and these epistemic rules are far from comprehensive) are those that cover parsimony or simplicity. Parsimony can guide us in justifying a belief in a particular theory to explain a particular

phenomenon, but *ceteris* is rarely *paribus* in these cases. It is rare to find a case where two theories are equal in explanatory power but only differ in their internal complexity. Rather, choices between theories can often hinge on much larger considerations of simplicity: which theory results in a simpler overall conceptual scheme. Quine mentions this as an important motivation for belief revision; in speaking of ontological questions (but could be applied more broadly), he says that “conservatism figures in such choices, and so does the quest for simplicity.”⁵

The problem here is articulating what exactly it is for one conceptual scheme or set of beliefs to be simpler than another. Perhaps simplicity is a matter of minimizing the types of entities that one believes in (allowing for relations such as composition); in this case, dualism is preferred over pluralism, monism over dualism, etc. Alternatively, simplicity may be a matter of minimizing the number of beliefs one has while still being able to sufficiently explain and predict things to the satisfaction of the individual. These two types of simplicity are applied via global epistemic rules, but perhaps the simplicity that some individuals seek is a type that is applied only locally, namely, in cases of belief revision. For example, when an individual is revising a belief in the face of new experiences, one consideration can be that which would cause the fewest overall changes in the individual’s current set of beliefs. Thus, a preference for simplicity might be a preference for minimizing disruption. *Ceteris paribus*, the revision in beliefs that causes the fewest changes in other beliefs is to be preferred.

⁵ Quine 1953, p. 46

We now have five different types of epistemic rules to keep in mind when approaching the debate over physicalism, but the two that will be a main focus are those governing explanation and parsimony. In short, to what extent should we complicate or disrupt our conceptual schemes to be able to explain some particular phenomena in the cognitive sciences? Prediction is not the problem, for everyone will agree that the models discussed below do indeed have good predictive power. Rather, to what extent should we accept that the models and theories are correctly describing objects and their causal powers, thus entailing that there actually are such objects with such causal powers? And can we resolve this question without appealing to epistemic rules regarding explanation and simplicity?

It is this framework of epistemic rules, and the tension between explanation and simplicity in particular, that will frame the rest of the discussion on physicalism. Before starting in on that debate, however, a few general working hypotheses must be stated. In what follows, we will assume that a few epistemic rules are shared amongst the discussants, although how much of a priority they put on these rules will vary. First, for phenomena for which the locus of prediction and/or explanation is far from direct experience, the evidential ideal is the scientific method (broadly construed). That is, for many everyday beliefs, direct experience is justification enough: nothing more need justify that I am typing on a keyboard than my current experience that I am indeed typing on a keyboard. In contrast, for complex systems of phenomena that are not amenable to direct perception or simple intervention, the scientific method is the peak of

justification and will trump conclusions produced by other means. By the scientific method, I mean something resembling an experimental method with proper controls (e.g., randomization for statistical populations). This basically comes down to the simple commitment that if you want to know how the world works, you have to do science. I take this to be uncontroversial and widely shared among the discussants. Given this assumption, part of the debate over physicalism will be ignored: that of *a priori* physicalism. The debate over *a priori* vs. *a posteriori* ontologies is a complicated one that will not be gone into here. Rather, we will simply be examining the debate over *a posteriori* physicalism; that is, ontological commitments (and their attendant theses, such as physicalism) cannot be justified without appealing (in part or in whole) to experience. Because the scientific method appeals to experience and because we have earlier linked ontological commitments with explanation, and explanation with the scientific method, we can see that the epistemic rules so far articulated are compatible with *a posteriori* physicalism (whether they are incompatible with *a priori* physicalism is a separate question).

A second related working hypothesis is that being committed to a theory that invokes certain causal powers is a reason to believe in the existence of objects that have those causal powers. This is just another way of restating a point made earlier: successful explanations can be used to justify ontological claims, but whereas earlier it was pointed out to be a result of making explanations, here we assume that individuals not only make these inferences from explanation to ontology, but also are justified in

making these inferences. This, of course, doesn't assume that successful explanation is always sufficient to make ontological commitments; this will clearly depend on the other epistemic rules that an individual holds and the prioritization thereof.

To sum up: we have outlined a way to approach philosophical debates (with the debate over physicalism as our ultimate goal) that is committed to examining not only the specific arguments that the discussants put forth but also to examining the epistemic context in which each of the discussants make their arguments, that is, what epistemic rules each discussant holds and how they prioritize their epistemic rules. Ultimately, we may find that the disagreement between the discussants is not because they hold different views on the validity or soundness of an argument, but because they have different epistemic rules or priorities.

2. Two Questions of Physicalism

Finally, then, to physicalism itself. Stoljar provides a clear framework for discussing physicalism. As a first step, he distinguishes two questions that must be answered by any discussant in the physicalism debate. First is what Stoljar calls the “interpretation question”: “What does it *mean* to say that everything is physical?” Second is what Stoljar calls the “truth question”: “Is it *true* to say that everything is physical?”¹ So, what is physicalism, and is it true?

The interpretation is very tricky, and enormous amounts of ink have already been spilled in trying to answer it. In particular, much has been written trying to parse “everything” in “what does it mean to say that everything is physical,” which Stoljar calls the “completeness condition”. For the last forty-odd years, the most common interpretation of “everything” has involved supervenience. Take Lewis’s definition of materialism as representative: “Among worlds where no natural properties alien to our world are instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates.”² We need not go into what exactly a natural property is in Lewis’s sense, but Lewis’s definition of materialism does state the core proposition from which many debates on physicalism occur: physicalism is true for any world w iff a complete physical duplicate of world w is also a complete duplicate in every other respect. That is, if we set about only duplicating the world w according to its physical properties and relations, then as a consequence we have duplicated world w in

¹ Stoljar 2009

² Lewis 1983, p. 363

every respect. This thesis is stated in shorter terms by saying that all properties are either physical properties or supervene on physical properties.

Something like this supervenience thesis is shared by most characterizations of physicalism, and because of this, it is often characterized as “minimal physicalism”. However, these various characterizations of physicalism can vary in what other requirements or commitments are brought along. For example, one stripe of eliminativist might argue that there are no such things as mental properties and that the only properties that exist are physical properties. This is compatible with the supervenience thesis, as the latter states that if there are any other properties besides mental properties, then they supervene on the physical. Thus, minimal physicalism as characterized by the supervenience thesis is noncommittal on the existence of non-mental properties.

There are many complicated objections to this characterization of minimal physicalism, challenging that it either is too inclusive or too exclusive to intuitively fit with what we call “physicalism”. These examples and subsequent modifications to the thesis of minimal physicalism are too numerous and esoteric to enumerate, but one example from Horgan gives a good flavor of the debates. Consider, Horgan says, a world that is identical to ours in every physical respect but also contains Cartesian souls and other spiritual substances that do not causally interact with any of the physical objects. If physicalism were true at our world, then such a spirit-filled world would be impossible. Yet surely we do not want the thesis of physicalism to rule the mere

possibility of this type of world. Likewise, if such a world is possible, then physicalism is false at our world even if our world only contains physical properties. That the possibility of such a world could render physicalism false at a world that has only physical properties also seems intuitively wrong.³ The point here is not to go into such a debate, but to point out how a large part of the debate over physicalism (particularly the interpretation problem) involves honing precise characterizations of supervenience that will not conflict with our intuitions about the consequences of physicalism, such as how various modal truths affect and are affected by the truth of physicalism in our world.

Those more concerned with the truth question (such as myself) need not be committed to any answer to the interpretation question. To circumvent the interpretation question, we can ask a question of minimal physicalism itself: suppose minimal physicalism or one of its refined descendents is the correct characterization of what it means for *everything* to be physical (ignoring, for the moment, what it means for something to be *physical*); what kind of evidence can we gather in favor of such a hypothesis? Whether one is a realist or not about possible worlds, we clearly cannot examine possible worlds that our like our world in every physical respect. Because we cannot do so, we must reframe the characterization of physicalism in modal terms: physicalism is true for any world w iff it is impossible for a complete physical duplicate of world w to differ from world w in any respect. We introduce modal terms precisely because we cannot examine physical duplicates of our world, but we do have established procedures for gathering evidence of some types of modal claims. The

³ Horgan 1982

procedure to justify claims of nomological possibility, for example, is the scientific method (again, broadly construed). Take the ideal gas law: it allows us to justify claims of what the temperature of a certain amount of gas would be if we varied its pressure or volume within certain ranges. Because the ideal gas law ignores some properties of gases (hence, "ideal gas"), there will be error ranges on its predictions. However, it can still be used to justify counterfactual or modal claims. The statement "It is impossible for a noble gas to increase in temperature while also decreasing in volume and pressure for ranges x , y , z of temperature, volume, and pressure respectively" can be justified by appeal to the ideal gas law, which itself has been established via the scientific method (theoretical and conceptual innovation, experimental confirmation, etc.). Thus, if the type of possibility in our re-characterization of minimal physicalism is nomological possibility, then we do have some way to gather evidence for it: look at what the science says. If science ranges over non-physical properties that vary independent of any physical properties, then clearly minimal physicalism is false, as it would be nomologically possible for there to be a world that was identical in all physical respects but differed in some other aspect.

Unfortunately, even if all of current science supported the supervenience on physical properties of all properties invoked in scientific theories, that may not be enough to provide support that physicalism is true. Where above we interpreted the type of possibility in the re-characterization of minimal physicalism as nomological possibility, that may be too weak. Of course physicalism is false if science endorses non-

physical properties, but that's not what is of interest: rather, what physicalism is concerned with is metaphysical possibility. Indeed, natural laws, which are the source of justification for claims about nomological possibility, are included in what is meant by "a complete physical duplicate". Examining nomological possibility, then, doesn't do anything because the presumption is that science only ranges over physical entities, and thus all nomological claims true in the actual world are also true in a world that is a complete physical duplicate. What is at question is whether it is *metaphysically* possible that such a duplicate may differ from the original world.

Metaphysical possibility is not an uncontroversial (or even particularly clear) notion. Whether one gathers evidence for metaphysical possibility via conceptual analysis or empirical investigation or whether metaphysical possibility is even a coherent notion are not debates that I will engage in. Instead, note that any investigation of metaphysical possibility and its relation to physicalism presumes that in the current world, laws only govern physical properties: physicalism is true for any world w iff it is *metaphysically* impossible for a complete physical duplicate of world w to differ from world w in any respect. If science ranges over non-physical entities and non-physical laws, then the thesis is clearly false, but if science does not, the thesis is not obviously true. In short, the justification of this thesis assumes that the weaker version that uses nomological possibility instead of metaphysical possibility has not been disproven, otherwise there would be no need to strengthen the thesis to the version invoking metaphysical possibility. Thus, we will leave behind the main debate over the

interpretation question for now, as even if we did settle on a characterization of supervenience physicalism that met all of the esoteric objections, it's not clear how we could justify a position that involves metaphysical possibility rather than just nomological possibility.

3. Physicalism as an Empirical Prediction

In place of a type of supervenience physicalism that relies on specifically metaphysical concepts, we must cast about for a different characterization of physicalism for which evidence can be marshaled (and thus is not trivially true, e.g., to count as physical any type of object that scientific theories range over) but which is also not clearly false (e.g., Democritean atomism). One such characterization is given by David Papineau, who seems less concerned with possible worlds and more concerned with a characterization of physicalism for which some time of empirical support can be marshaled. For Papineau, physicalism is the conjunction of two theses: the supervenience of all non-physical properties on physical properties and the token congruence of all non-physical events with physical events.¹ Papineau's invocation of physicalism need not send us running to the hills for fear of more abstruse metaphysical debate. Rather, it is a straightforward thesis about the world that we live in: "two systems cannot differ chemically, or biologically, or psychologically, or whatever, without differing physically; or, to put it the other way round, if two systems are physically identical, then they must also be chemically identical."² Supervenience is not sufficient for physicalism, although it is necessary, as it is compatible with our world being clearly non-physical. For example, in our world, mental events might be epiphenomena: they are caused by physical events but have no causal powers themselves. If having no causal powers is not enough to disqualify mental events from

¹ Papineau 1993, Ch. 1

² Papineau 1993, p. 10

being physical, then also imagine that they have no spatial or temporal location, mass, volume, etc. In this type of world, mental events would supervene on physical events, but the mental events are clearly not physical even in the loosest sense of the word. Thus, physicalism would be false.

Because of this, Papineau adds the condition of token congruence, which requires that all events characterized as non-physical be identical with physical events ("characterized" being used to avoid such contradictions as essentially non-physical events being identical with essentially physical events; clearly, if physicalism is true, then there are no essentially non-physical events). The token identity of non-physical and physical events rules out cases such as protoplasmic epiphenomenalism. Token congruence is not itself sufficient for a reasonable type of physicalism because there are cases that are token congruent but not locally supervenient. For example, if consider broad representational content to be legitimate, then a broad mental representation will not supervene merely on the brain state of the individual, but on relations, history, etc. Broad content may supervene in a non-local sense, but allowing non-local supervenience can makes empirical verification difficult.³ The less local the supervenience is, the more difficult it becomes to gather evidence for; the limiting case is global supervenience, for which the complete set of non-physical properties and relations supervenes on the complete set of physical properties and relations, and no further specification is possible. Gathering evidence for global supervenience of this kind is practically impossible.

³ Note that Papineau is not arguing against broad content, but just using it as an example of relatively non-local supervenience.

So how can we gather evidence for the conjunction of the local supervenience of the non-physical on the physical as well as the token congruence of the non-physical with the physical? One option is to do an exhaustive investigation of the whole of science. Are there theories for which either local supervenience does not hold or for which token congruence does not hold? One counterexample is enough to disprove physicalism, as physicalism is a thesis that states that *all* non-physical properties are supervenient on the physical and that *all* non-physical events are congruent with the physical. Unfortunately, figuring out just what a counterexample would look like is not easy. We'll go into this more below, but first let's look at Papineau's approach. Instead of searching for counterexamples, Papineau takes an inductive shortcut based on what he considers to be an empirically well-supported truth of physics. According to Papineau, physical systems are complete: any physical effects are fully fixed by their physical antecedents, and this fixation is a consequence of stable physical law (in a different world, perhaps physical effects are fully fixed by their physical antecedents, but in a non-systematic way). Thus, physical causes are fully sufficient for their physical effects. As a result, "if two systems are physically identical and in the same physical contexts, they will issue in the same physical consequences or chances thereof."⁴ Papineau's other premise requires that differences in the mental be physically manifestable in some manner. The most common way for mental differences to manifest are behavioral differences, but they may be manifested in other ways, such as in physical relations (e.g., to accommodate broad content). This premise does rule out epiphenomenalism, but

⁴ Papineau 1993, p. 17

barring some positive evidence for epiphenomenalism, we can accept this premise as reasonably well-supported empirically. The consequence of these two premises is that mental differences can only issue from physical differences. If two systems are mentally different, then by the second premise they result in different physical consequences. By the first premise, if there are different physical consequences, then there must have been different physical antecedents (by nomological necessity). Thus, the mental supervenes on the physical: there are no mental differences without physical differences, but there can be physical differences without mental differences.

Note that Papineau's argument only concerns mental properties and not all non-physical properties. To support physicalism, arguments of this type would have to be constructed for each type of non-physical property (social properties, economic properties, etc.). I will leave the plausibility of the manifestability of social and economic properties up to the reader. For now, let's assume that the manifestability of all non-physical properties can be supported on empirical grounds. Let's further assume that this argument is valid (as it seems to be). The main challenge to this argument is against the completeness of physical systems. The challenge to this completeness premise has been most thoroughly articulated by Tim Crane. Why, Crane asks, should someone who is already a dualist (or pluralist or what have you) believe in the completeness of physical systems? As it stands, it seems to simply beg the question against the non-physical having physical effects. So some type of evidence must be marshaled in favor of the completeness of physical systems. A dilemma emerges, however, when trying to

define what it is to be a physical system. Either 'physical' is defined according to current theories in the physical sciences, or it is defined according to future theories in some ideal finished physical science. If the former, the completeness of physical systems is open to empirical challenge; at the very least, much empirical legwork needs to be done to show that given the current status of the physical sciences, we are justified in believing in the completeness of physical systems. If the latter, then the completeness of physical systems is a trivial thesis, as it seems possible that future physical theories may range over completely different objects than current theories; compare, for example, the objects over which Newtonian physics ranges and the objects over which quantum mechanics ranges.⁵ The conceptual distance between the two sets of theories is enormous. The conceptual distance between current physical theories and future physical theories may be equally large. Given this, it seems possible that future physical theories may range over objects or properties that are currently characterized in non-physical terms (such as mental properties).

Because of the near impossibility of predicting the future state of science, one would expect Papineau to grapple with the first horn of the dilemma and try to marshal evidence for the completeness of physical systems according to the current state of physical science. This is a daunting task, surely, and the current state of physics (or any science) is not expected to remain as it currently is indefinitely. However, the fact that science progresses and that the best theories today may be rejected in the future doesn't seem to be a sufficient reason for rejecting ontological conclusions drawn from those

⁵ Crane 1991

theories. If we recognize all such ontological claims as defeasible, then we are still justified in making such ontological inferences from theoretical success. To deny this because new evidence may be introduced or newer, more explanatory theories may be articulated in the future has enormous consequences: as a result, Newtonians would not have been justified in believing in the gravitational force as articulated in Newtonian physics, Rutherford would not have been justified in believing in an atomic nucleus, and so forth for countless other ontological claims made as a result of the explanatory power of a particular theory. If ontological claims over and above those drawn from direct experience (e.g., trees exist) are to be justified, inferring them from successful and experimentally tested theories seems like the strongest possible way to justify them. Thus, if successful theories can't justify ontological claims, then nothing beyond basic experience can. This would leave us with a greatly simplified ontology, one that accepts only mid-sized objects that can be directly confronted by the senses; theories (and science as a project) may then simply be viewed as tools for achieving particular outcomes. This may be how some philosophers think of science (and ontological claims more broadly), but this does not seem to be the general perspective of those engaged in the debate over physicalism. Rather, the question of physicalism is if those many theoretical entities invoked by the special sciences (read: anything that's not physics or chemistry) are really, in some sense, ultimately physical.

Recognizing the defeasibility of ontological claims drawn from successful theories but still being justified in making such claims is really just a specific application

of a general epistemic practice: all claims we make are defeasible because for any claim, we could be confronted by new evidence that would cause us to modify or reject that claim. If non-defeasibility is required to truly be justified in making a claim, then there will be very little that one could actually be justified in believing. The argument that physicalism cannot be justified by looking at the current science because the current science could be wrong assumes such a non-defeasibility requirement. Yet as we've seen, unless an individual has very atypical preferences, it is very unlikely that anyone (much less anyone engaged in the debate over physicalism) actually holds such a non-defeasibility requirement for the justification of belief.

The first horn of Crane's dilemma, then, has not been eliminated, but it at least has been defanged. Looking at the current science is a viable option for justifying the completeness of physical systems. It is surprising, then, that Papineau opts to grapple with the second horn of the dilemma. Papineau does admit that it is possible that future theories will range over non-physical objects and properties, but he does not think it is very likely: "It seems to me highly unlikely that the psychological will turn out to be a part of the physical...The history of science yields a great deal of empirical evidence about the *kind* of causes that are responsible for the motion of stones and other kinds of matter...[This evidence] does, it seems to me, provide sufficient ground for concluding that mental categories are not among [those causes]."⁶ Given this pronouncement we must ask what kind of evidence can justify claims about what objects or properties future scientific theories will range over. Papineau mentions the history of science, so it

⁶ Papineau 1993, p. 31

seems that he is making a kind of historical induction: theories in the past have not ranged over non-physical objects or properties; therefore, theories in the future will not range over non-physical objects or properties. Of course, if this is what he means, this is obviously false. There are lots of theories in the past and present that range over objects that aren't essentially physical; psychological theories in the past have ranged over experiences, and current psychological theories range over things like representations, memories, perceptions, etc. Economic theories range over preferences, utility, interest rates, and aggregate demand, among other things. None of these are essentially physical (that is, they aren't characterized in the terms of physics or chemistry); if they were, then there would be no debate over physicalism.

Therefore, we must reformulate the induction: physical theories in the past have not ranged over non-physical objects or properties; therefore, physical theories in the future will not range over non-physical objects or properties. But this will not do. The first part is a definition rather than an observation. A more specific statement must be made: physical theories in the past have not ranged over objects or properties of types X, Y, and Z; therefore, physical theories in the future will not range over objects or properties of types X, Y, and Z. Here, X, Y, and Z represent specific objects and properties that are characterized in ways that are not essentially physical (note that this is not the same thing as essentially non-physical), such as experiences, memories, perceptions, interest rates, preferences, and so on. Unfortunately, this characterization still doesn't fully capture the type of induction that would be needed to justify the

completeness of physical systems without appealing to the truth of current theories and without begging the question against those who think the completeness of physical systems is false given current theories. After all, the reason why Papineau seems loathe to use the current state of physical theories to justify the completeness of physical systems is that there are phenomena that are unexplained by current physical theory. These explanatory gaps may be used to justify rejecting the completeness of physical systems. At the very least, given the current state of the physical sciences, it is not obvious that physical systems are complete. Thus, we must again modify the induction: physical theories in the past have not ranged over objects or properties of types X, Y, and Z; therefore, physical theories in the future will not range over objects or properties of types X, Y, and Z and physical theories in the future will explain the phenomena that theories that range over X, Y, and Z currently explain. As a basic induction, this fails. There is nothing in the fact that past physical theories have not ranged over X, Y, and Z to suggest that they will explain the phenomena that theories that do range over X, Y, and Z explain. The first part of the induction needs to be bolstered.

One final modification: physical theories in the past have not ranged over objects or properties of types X, Y, and Z and the explanatory range of the physical sciences has been increasing throughout their history; therefore, physical theories in the future will not range over objects or properties of types X, Y, and Z, and the theories of the physical sciences in the future will explain the phenomena that theories that range over X, Y, and Z currently explain. The implication of the latter half is that future physical theories will

explain the phenomena of the special sciences (those that range over X, Y, and Z) without themselves ranging over X, Y, and Z. This induction is really two inductions; the first half of the antecedent only supports the first half of the consequent, and the second half of the antecedent only supports the second half of the consequent. Thus, the first induction is identical to an earlier form of the induction above; Induction 1: physical theories in the past have not ranged over objects or properties of types X, Y, and Z; therefore, physical theories in the future will not range over objects or properties of types X, Y, and Z. Induction 2: the explanatory range of the physical sciences has been increasing throughout their history; therefore the theories of the physical sciences in the future will explain the phenomena that theories that range over X, Y, and Z currently explain.

Both of these inductions are needed to justify physicalism in the form of Papineau's prediction: future physical theories must explain most of what can't be explained now and do so with a particular set of theoretical resources. Papineau admits that the current categories of physics ("the categories of energy, field, and spacetime structure"⁷) may not be up to the job. Hence, the need for these inductions, but do we have any reasons for believing that these inductions are even minimally acceptable (much less being strong inductions)? Whether the inductions will be acceptable are in part up to the individual: some people may have higher inductive standards than others (and having high inductive standards isn't necessarily a good thing). What we can do is compare them to two other well-established inductive standards: that of everyday

⁷ Papineau 1991, p. 38

experience and that of the scientific method. In everyday experience, we make basic inductions: I have eaten at this restaurant twice now and haven't liked what I've had; therefore, I will not like the food here in the future. That is not a rigorous induction by the standards of science, but is one that most people would accept as reasonable. We need not have high inductive standards for deciding what restaurants we like. In fact, we would likely think a person unreasonable if they had everything off the menu at least once before deciding if they like the food at a particular restaurant. In contrast, the scientific method is more rigorous. Experiments are run with controls for variables that may interfere with causal relation we are trying to assess. In medical trials, for example, randomizing the population can help control for sampling biases, and double-blinding can help control for various confirmation biases.

Are the inductions supporting physicalism like either of these types of inductions (and thus should be held to that particular low or high standard)? Earlier we assumed that ontological conclusions can be drawn from the best scientific theories, and presumably those scientific theories are established via something approximating the scientific method. If physicalism is the thesis that currently we are justified in believing that the world consists of only physical objects because Inductions 1 and 2 hold, then it seems that we would want to hold those inductions to a fairly high standard, perhaps approximating the scientific method. Although we do use everyday experience to make inferences concerning the existence of particulars (there is a car parked across the street because I am currently looking at it), it is rarely used to make inferences concerning the

existence of types of things, particularly types of things that are novel or complicated (by “types”, I mean an ontological category as broad and as basic as possible; thus, a new type of propulsion system for boats, for example, would not count). Indeed, people are often admonished for making ontological conclusions based on basic experiences: you shouldn’t have concluded that it was a ghost based on your experience alone, as the more thorough examination shows it to be a trick of light and a strange echo. Thus, *prima facie*, our inductive standard for ontological conclusions should be fairly high. We might not require a perfectly controlled experiment for all ontological inferences, but the evidence must at least somewhat approximate a controlled experiment.

Unfortunately, it seems practically (if not conceptually) impossible to gather that kind of evidence for Inductions 1 and 2 since the inductions are historical inductions. What the inductions suppose is that there is a relationship between some properties of theories in the past, and these properties of the theories can be used to predict the future state of theories. We can’t run experiments on these theories for a variety of practical and conceptual reasons, the first simply being how theories can be used as an object of investigation. If we had unlimited power and resources, what would an experiment on these theories look like? Other types of historical propositions are at least amenable to this type of experimental investigation. For example, one might assert (simplistically) that industrial revolutions are a necessary and sufficient condition for the rise of social democracy. There is no way one could actually run an experiment to test this proposition, but we can at least conceive of an experiment to test it: set up several states

with starting conditions that will ensure an industrial revolution in some and no industrial revolution in others, and then see how many develop social democracy. The difference with the scientific theory case is that instead of using an induction (or theoretical experiment) to justify a truth as in the social democracy case, the scientific theory case would be using an induction to infer what propositions would be justified by some other criterion *but still in line with the induction*. If the prediction of physicalism actually comes to pass, then physicalism would be justified by the current science, not an induction based on the history of science. The induction thus bypasses the exact means by which such truths are generally justified. This does not necessarily make the induction weak, but it should make us ask why we shouldn't use the well-established method of making ontological conclusions: constructing theories and rigorously testing them.

In the end, the inductions are not particularly strong inductions, similar to most historical inductions, and this one is more complicated than most. Whether the inductions indeed justify one believing in physicalism depends on the particular individual's epistemic standards, but it would be an odd set of standards to hold that the scientific method is the ideal for drawing ontological conclusions and then also accepting this particular induction supporting physicalism. Either the individual would, for non-epistemic reasons, lower his/her inductive standards in that particular case, or have such low inductive standards to result in a chaotic set of beliefs. Consider the equally weak induction: most scientific theories in the past have been wrong, therefore

current scientific theories are wrong. If one accepted this, then there would be no justification for believing any current scientific theory, and one would probably despair of ever accepting any scientific theory. The inductive strength of this particular induction seems to be about on par (insofar as we can judge these things) with the induction supporting physicalism. Thus, to have such low inductive standards is to result in either a rejection of all current scientific theories or a set of contradictory beliefs. Presumably physicalists would not want either outcome.

Ultimately, these inductions can be justified, but the inductive standards one would have to hold to justify them would lead to consequences that are likely undesirable. Remember, however, that these inductions are the result of trying to grapple with the second horn of Crane's dilemma: defining what is physical in terms of future theories. There is still the first horn left for physicalists, and though daunting, it may be a better way to justify physicalism: given the current science, we are justified in being physicalists. For this approach to physicalism, we can draw off of the work of John Bickle.

4. Bickle's Physicalism as a Non-inductive Project

It must be noted at the outset that, in fact, Bickle's work is another attempt at trying to predict what future theories will be like rather than trying to justify physicalism via current theories: "Physicalism is the *prediction* that theories from intentional psychology ultimately will reduce to theories pitched at the level of physical mechanisms..."¹ However, unlike Papineau, the evidential basis for this prediction is much more narrow. Bickle is not drawing inductions across the history of the sciences, but rather making predictions based on the current intertheoretic relations between some theories in psychology and neuroscience. Physicalism in this sense is a project that must make local inductions based off of current intertheoretic relations to future ontologies. One of the main cases for Bickle is the relation between cognitive theories in associative learning and the neurobiological theory of long-term potentiation (LTP).² The main argument seems to be this: there is a reductive relation between associative learning and LTP. Because associative learning is reduced to LTP, then the ontological resources of LTP suffice for explaining the phenomena involved in associative learning. Furthermore, the ontological resources of LTP are limited to the strictly physical. Given this reductive case and a few others (such as memory consolidation being reduce to molecular mechanisms in the hippocampus), we are justified in predicting that theories in psychology (current and future) will be reduced to neuroscientific theories. Therefore, we are justified in believing that the ontological resources provided by the physical

¹ Bickle 1998, p. 17

² Bickle 1998, Ch. 5

sciences will be sufficient to explain the phenomena explained by (current and future) psychological theories. Therefore (finally), we are justified in believing that there are no psychological objects that are non-physical.

Unfortunately, this prediction runs afoul of the problems that plagued the more expansive inductions of the previous section. First, it is an induction that states we are justified in believing in a particular ontology *because* we are justified in believing that there will a theory in the future that justifies believing in that particular ontology. If that's not enough to make you uneasy, the inductive standards to accept the induction still have to be either arbitrary (accepting this and nothing else) or somewhat low, as this is nowhere near the strength of typical scientific inductions (even though it may be stronger than Inductions 1 and 2 above).

To avoid these problems, Bickle's project can be viewed in a different light. Instead of using the specific cases of reduction to make a prediction about future theories, we can view the cases as steps in a piecemeal justification of physicalism using current scientific theories. Thus, via Bickle, we can try and tackle the first horn of Crane's dilemma rather than the second. To do so, let's look at one of Bickle's prime cases: the reduction of associative learning to LTP.

Pavlov's dogs have worked their way into the cultural consciousness even if "classical conditioning" is still a term mostly used in introductory psychology classes. Since Pavlov's initial characterization of classical conditioning, countless different experimental procedures, behavioral anomalies, and mathematical models have been

created. The most famous model, and the one that Bickle uses as an illustration of the reduction of a cognitive theory to a neurobiological theory, is the Rescorla-Wagner model. Before diving into the model, we need briefly elaborate the experimental and theoretical context of the model. We will use the Pavlovian case of a dog salivating for simplicity even though most modern research on classical conditioning is done using a rabbit's nictitating membrane reflex or a rat's startle reflex. First, there are two types of stimuli: conditioned stimuli (CS) and unconditioned stimuli (US). Unconditioned stimuli are typically of some sort of biological value, such as food (positive value) or sounds representing danger (negative value). Conditioned stimuli can be practically anything: lights, tones, smells, the appearances of particular objects, and so on. In a particular experimental trial in which we want the animal to learn an association between a CS and a US, the CS is presented some time before the US (its presence can persist into the presentation of the US, such as if the presentation of a shape to an animal occurred during every feeding). In our example, the CS is a bell and the US is the presentation of food. Over several trials of a bell preceding the presentation of food, the dog will start to anticipate the food, and this anticipation comes in the form of the bell producing a response (the conditioned response, CR) that is normally produced by the food (the unconditioned response, UR); in this case, salivating. When the dog salivates in response to the sound of the bell, it has learned to associate the bell with the presentation of food.

There are many variables in this type of specific learning procedure: when the CS is presented in relation to the US, the intensities of the CS and the US, the number of

different CSs, alternative trials with different CSs, alternating trials with and without the US, and so on. The main variant and experimental phenomenon we need to introduce to understand the explanatory power of the Rescorla-Wagner model is the learning phenomenon known as 'blocking'. Imagine that we have a dog that is sufficiently conditioned to the bell: when the bell rings, the dog salivates a good amount. Now imagine that we start a new block of learning trials in which the bell rings at the same time that a light flashes, after which food is presented. Will the dog learn to associate the light with food? After all, like the bell, it is presented before each presentation of food and only before each presentation of food (unlike, say, the door of its cage, which is ever-present and thus does not become associated with food). In this case, if the dog light is flashed by itself, the dog will salivate very little (if at all), and, more importantly, much less than a dog that was trained with the bell/light combo from the very beginning (i.e., one that was not already trained to the bell). In this case, because the bell was already a signal for the food, the light presented no new information. The dog was already anticipating food. The bell 'blocked' the light.

Now that we have a basic grasp of classical conditioning and blocking, we can introduce the Rescorla-Wagner model³:

$$\Delta V_x = \alpha\beta(\lambda - \Sigma V)$$

V is the strength of the association between two stimuli, and λ is the strength of the US presented in a particular trial (α and β are unimportant parameters for our purposes, but

³ Rescorla & Wagner 1972

we will arbitrarily set their product to be 0.5). This equation models how much a particular association between stimuli will change after one experimental trial. ΣV is the sum of all activated associations (i.e., those associated with the particular US presented). Thus, the equation states that a particular association will increase in strength equal to the amount of the intensity of the US minus the strengths of all associations active upon presentation of the US. This allows for basic associative learning. For the first time the bell is rung and food is presented, there are no stimuli already associated with food that have been presented; thus, there are no active associations: ΣV equals zero. Because food was indeed presented, λ is positive; thus, ΔV_{bell} is non-zero and positive. If λ is equal to one and $\alpha\beta$ equal to 0.5, then ΔV_{bell} equals 0.5. For the next learning trial, ΣV is no longer zero (because ΔV_{bell} is 0.5). Thus, on the second trial, we have $0.5(1 - 0.5)$, which equals 0.25. That's the change in the association of the bell to the food for the second trial. The bell becomes more associated with the food, but by less than after the first trial. If we keep repeating the learning trials, we will see that learning is asymptotic: ΔV_{bell} will increase at a slower and slower rate, never quite reaching a value of one.

Now suppose we have a dog that has been through many trials and now has a V_{bell} equal to 0.99. If, in a new trial, we shine a light in addition to ringing the bell, then the light will gain very little association: $0.5(1 - 0.99)$. ΔV_{light} would only equal 0.005! This is much less than the initial learning trial with the bell when after just one trial, V_{bell} equaled 0.5. In this case, the bell has blocked the forming of an association between the light and the food.

The ability of the Rescorla-Wagner model to explain blocking as well as other features of classical conditioning (such as the asymptotic nature of learning) garnered it much acclaim, and rightfully so. Bickle uses this theory as the psychological theory to be reduced to a neurobiological theory, in particular, long-term potentiation (LTP). In very coarse terms, the functionality of LTP can be captured by the slogan “Fire together, wire together”. We need not get into all the biological details of LTP, but the basic idea is that when two neurons that are connected via synapses fire at the same time (or within a particular time interval), the connection between the two is strengthened (either by the creation of more synapses or the subsequent release of greater amounts of neurotransmitter from existing synapses).⁴ To connect it to classical conditioning, we can conceive of the downstream neuron as representing the US and the upstream neuron as representing the CS. In addition, like associative learning, the connection between the two neurons has a strength ceiling: if the firing of one neuron causes the neuron onto which it has synapsed to fire at its maximum rate, then the connection between the two is at its maximum strength. This captures the basic paradigm of classical conditioning. A neurobiological story of learning trials might go as follows: when the bell is rung, the “bell neuron” fires. At first, the bell neuron only has a few synapses on the “food neuron”, and the synapses don’t release very much neurotransmitter when the bell neuron fires. Thus, when a bell occurs and the bell neuron fires, the food neuron is only marginally excited and its firing rate is not increased. However, in the experimental trial, the bell neuron fires when the bell is rung, and then the food neuron fires at the

⁴ For a reviews of the LTP literature, see Malenka & Nicoll 1999 and Lisman 2003

presentation of the food. Because the two fire so closely together in time, the connection between the two is strengthened (suppose that the bell neuron grows more synapses that connect to the food neuron). Thus, the next time the bell neuron fires, the food neuron's firing rate is increased somewhat. After repeating these learning trials many times, the food neuron's firing rate is increased to its maximum each time the bell neuron fires because a great number of synapses have formed between the two.

LTP thus maps nicely onto basic classical conditioning, but what about blocking? Isolated to one or two neurons synapsing onto another neuron, LTP cannot explain blocking. If the bell neuron is already fully trained (i.e., its firing causes the firing rate of the food neuron to max out), the firing of a "light neuron" that synapses onto the food neuron would still cause it to increase its synaptic strength. LTP of only excitatory neurons can't enact $(\lambda - \Sigma V)$. However, if we introduce inhibitory neurons and networks of multiple neurons, then it is possible to produce blocking effects.⁵ The details are unimportant for our purposes, but presuming that networks of neurons that change according to the rules of LTP can explain the behavioral consequences of both standard learning trials and blocking, then it seems we have a potential case for theoretical reduction or elimination. Whether it is reduction or elimination depends on how similar the workings of the two theories are. If, for example, blocking is neurobiologically produced not via an explicit summation mechanism but by some other mechanism, then this would be more of an elimination than a reduction. Its status as an elimination or reduction, however, does not matter for the purposes of Bickle's project. Rather, what

⁵ See Hawkins and Kandel 1984

this example purports to do is to show that we don't need to invoke any entities over and above the neurobiological. In this case, there are no associations or representations over and above the neurons and their firing rates.

Unfortunately, even this basic case in the project of physicalism is not so clearly supported. The Rescorla-Wagner model was powerful compared to contemporary models when first articulated in 1972, but since then, problems have accumulated and new models have been proposed. For example, consider an experiment in which both experimental and control groups were trained to associate a bell with a small food reward. Next, the experimental group was presented with trials in which the bell and a light preceded the small food reward; the control group received no trials. Finally, both groups underwent trials in which the bell and the light preceded a larger food reward (in terms of the Rescorla-Wagner model, λ was much larger in these final trials than the earlier trials). How does each group respond to an isolated presentation of the light? According to the Rescorla-Wagner model, there should be no difference because for the experimental group, the bell should have blocked the light in the second set of trials; both groups should have some response to the light because it was trained in conjunction with the bell on the stronger stimulus (in these trials, the bell didn't predict such a large λ because they were trained to the asymptote of the smaller λ in the initial trials). But when this experiment is actually performed, the experimental group actually responds *less* to the light than the control group.⁶ It's as if in the trials in which the light

⁶ Mackintosh & Turner 1971

was blocked by the bell, the animals learned to ignore the light and had to overcome that in the final trials.

In the wake of this and other problems with the Rescorla-Wagner model, new models were developed that could account for these findings.⁷ The explanatory scope of these models is much wider than the Rescorla-Wagner model, and likely much wider than current models of LTP. Thus, we have one theory (LTP) that explains particular findings in neurobiology (e.g., changes in synaptic density) and looks like a possible explanation for some basic learning behaviors.⁸ We also have another theory (one of the contemporary learning models) that explains quite a bit of behavior, much more than LTP may explain. What ontological conclusions can we draw from this situation? This is where conclusions may differ depending on one's epistemic rules and prioritization thereof. If explanation is favored, then one can justifiably conclude that the entities of the contemporary learning model exist, and these entities are non-physical, as they have not been reduced to, identified with, or eliminated in favor of physical entities. If conceptual simplicity is favored, then perhaps such ontological conclusions would not be justified, as they do complicate things: they entail the causal interaction of the physical with the non-physical. If LTP does indeed explain the basic learning scenario, then these non-physical entities may overdetermine the case of basic learning (as the contemporary models, of course, explain basic learning trials as well as the more complicated scenarios). Overdetermination is not messy in itself, but systematic overdetermination of

⁷ For two different contemporary approaches, see Brandon & Wagner 2002 and Pearce 2002

⁸ There is good evidence that LTP correlates with some types of learning, but it's not clear if it is the mechanism of learning; see Whitlock et al. 2006 for an example.

this kind has been a motivation for some to try to find a physicalist alternative. At the very least, accepting the ontological implications of these contemporary learning models introduces new categories of entities if one's previous ontology was purely physical.

In the end, the answer to the truth question of physicalism is "it depends on your epistemic rules". Physicalism as a *prediction* or *induction* is not justified unless one has extremely idiosyncratic standards of inductive evidence. However, physicalism as a thesis supported by current theories may be justified if simplicity is valued much more highly than explanation. Otherwise, physicalism as articulated above cannot be currently justified.

Bibliography

- Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: The MIT Press
- Brandon, S.E., & Wagner, A.R. (2002). Sometimes opponent process (SOP) model of conditioning. In J.H. Byrne, H. Eichenbaum, H. Roediger, III, and R.F. Thompson (Eds.), *Learning and Memory*, 2nd Edition. Farmington Hills, MI: Macmillan.
- Carnap, R. (1956). *Meaning and Necessity*. Chicago, IL: The University of Chicago Press.
- Clifford, W.K. (1877). The ethics of belief. *Contemporary Review*, 29, 289-309.
- Crane, T. (1991). Why indeed? Papineau on Supervenience. *Analysis*, 51:1, 32-37.
- Hawkins, R.D., & Kandel, E.R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, 91, 375-391.
- Horgan, T. (1982). Supervenience and microphysics. *Pacific Philosophical Quarterly*, 63, 29-43.
- James, W. (1911). *The Will to Believe and Other Essays in Popular Philosophy*. New York: Longmans, Green, and Co.
- Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61:4, 343-377.
- Lisman, J. (2003). Long-term potentiation: outstanding questions and attempted synthesis. *Philosophical Transactions of the Royal Society B*, 358, 829-842.
- Mackintosh, N.J., & Dickinson, A. (1971). Blocking as a function of novelty of CS and predictability of UCS. *Quarterly Journal of Experimental Psychology A*, 23, 359-366.
- Malenka, R.C., & Nicoll, R.A. (1999). Long-term potentiation—a decade of progress? *Science*, 285, 1870-1874.
- Papineau, D. (1993). *Philosophical Naturalism*. Oxford, UK: Blackwell Publishers.
- Papineau, D. (1991). The reason why: response to Crane. *Analysis*, 51:1, 37-40
- Pearce, J.M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning and Behavior*, 30, 73-95.

Quine, W.V.O. (1953). *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II*. New York: Appleton-Century-Crofts.

Stoljar, D. (2009). Physicalism. *Stanford Encyclopedia of Philosophy*. Retrieved February 10, 2012, from <http://plato.stanford.edu/entries/physicalism/>.

Whitlock, J.R., et al. (2006). Learning induces long-term potentiation in the hippocampus. *Science*, 313, 1093-1097.