sciendo

The Misuse and Failure of the Evolutionary Argument

Joseph Corabi
Saint Joseph's University

DOI: 10.2478/disp-2014-0013 BIBLID [0873-626X (2014) 39; pp. 199-227]

Abstract
The evolutionary argument is an argument against epiphenomenalism, designed to show that some mind-body theory that allows for the efficacy of qualia is true. First developed by Herbert Spencer and William James, the argument has gone through numerous incarnations and it has been criticized in a number of different ways. Yet many have found the criticisms of the argument in the literature unconvincing. Bearing this in mind, I examine two primary issues: first, whether the alleged insights employed in traditional versions of the argument have been correctly and consistently applied, and second, whether the alleged insights can withstand critical scrutiny. With respect to the first issue, I conclude that the proponents of the argument have tended to grossly oversimplify the considerations involved, incorrectly supposing that the evolutionary argument is properly conceived as a non-specific argument for the disjunction of physicalism and interactionist dualism and against epiphenomenalism. With respect to the second issue, I offer a new criticism that decisively refutes all arguments along the lines of the one I present. Finally, I draw positive lessons about the use of empirical considerations in debates over the mind-body problem.

Keywords
Mind-body problem, epiphenomenalism, evolutionary argument, William James, physicalism

Introduction

The evolutionary argument purports to be an argument against epiphenomenalism — the thesis that mental states and events have no causal effects.[1] The argument claims that epiphenomenalism can be

[1] Later, I will also make clear that I assume that epiphenomenalism is commit-

Disputatio, Vol. VI, No. 39, November 2014
Received: 26/08/2013 Revised: 08/11/2013 Accepted: 30/05/2014

disconfirmed on empirical grounds, rather than merely being counterintuitive. The evolutionary argument has a distinguished history, being introduced by Herbert Spencer (1871), and having been defended — in one form or another — by William James (1890), Karl Popper (Eccles and Popper 1977), and others. Here is a passage from James's classic statement of the argument:

> There is… [a] set of facts which seem explicable on the supposition that consciousness has causal efficacy. *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, exercise after rest, and a sound skin and unbroken bones at all times, 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. An animal that should take pleasure in a feeling of suffocation would, if that pleasure were efficacious enough to make him immerse his head in water, enjoy a longevity of four or five minutes. But if pleasures and pains have no efficacy, one does not see… why the most noxious acts, such as burning, might not give thrills of delight, and the most necessary ones, such as breathing, cause agony.[2]

While different figures have formulated the argument in subtly different ways, all of the ones following James's style have taken the central insight involved to be the basis for an argument for the causal efficacy of qualia; this central insight is that epiphenomenalism leaves the smooth correlation between negative qualia and harmful stimuli unexplained. Since all forms of interactionist dualism and virtually all forms of physicalism hold that qualia *are* causally efficacious, and all forms of epiphenomenalism hold that they are not, the argument is uniformly taken to be a non-specific argument for the disjunction of interactionist dualism and physicalism, and

ted to robust dualism of the sort proposed in Chalmers 1996. Serious defenders of epiphenomenalism have included Thomas Huxley (1874), Frank Jackson (1982), and William Robinson (2004).

[2] James (1890: 143-4), emphasis in original.

against epiphenomenalism.[3]

Although clever, the evolutionary argument has aroused its fair share of suspicion. The criticisms of it in the literature have been diverse, and also far from decisive. (See, for instance, Broad 1925, Jackson 1982, Van Rooijen 1987, Lindahl 1997, and Robinson 2003, 2007.)[4]

Given this background, my aim in this paper is twofold. First, I will show that the evidence the argument employs has been mishandled, even if we grant the important assumptions of the argument. Contrary to what its traditional proponents have led us to believe, it is not best conceived as a straightforward argument for the efficacy of qualia, and hence as a non-specific argument for the disjunction of interactionist dualism and physicalism. The matter is more subtle than this, and I will explain how the distinct kinds of evidence the argument employs pull us in a different direction from what someone like James supposed.

Second, once the traditional oversimplifications have been noted and an improved version formulated, I offer a new objection to the argument that decisively refutes it (or refutes it in anything like a traditional form, at least), by making clear once and for all the central mistake that plagues it. (The process of sorting out the earlier confusions will help to focus our efforts.) I will make the case that the central mistake lies in accepting one assumption in particular that is unjustified and almost certainly false. Although my primary aim

[3] I say that virtually all forms of physicalism hold this because there are a few physicalist views that hold that qualia are inefficacious because the neural states they supervene on (or are identical to) are physiologically cut off from the production of behavior. Such views are extremely rare and (relatedly) not usually considered plausible, so I ignore them here. Also, as I discuss below, for the purposes of this paper I do not classify as epiphenomenalist those physicalist theories that have trouble countenancing the causal efficacy of qualia for subtle metaphysical reasons (such as the ones that sometimes arise in connection with role functionalism) — I treat these as straightforwardly non-epiphenomenalist. Incidentally, I also assume throughout that all views must acknowledge the reality of qualia, even if they are ultimately reducible to or in some other metaphysically intimate way dependent on the physical. This is keeping with trends in the philosophy of mind over the past generation, where accounting for phenomenal consciousness has generally been considered of central importance.

[4] See also my response to Robinson (2007) in Corabi 2008.

here is critical, there are positive lessons to draw from this negative result; in particular, we gain some insight into what empirical considerations can genuinely help us to solve the mind-body problem.

I will begin by clarifying some important terms, and then formulate a canonical evolutionary argument against epiphenomenalism. I will then employ this canonical formulation to explore the ways in which the argument has mishandled the evidence, even conditional on the correctness of the assumption I will later subject to scrutiny. After providing an adjusted formulation that sidesteps these problems (though at the cost of complicating the commonly accepted conclusion of such arguments), I will then discuss the key assumption that drives these arguments toward their conclusion and explain why it drives them in this direction. The assumption is that physicalism, because it claims physical neural bases of qualia metaphysically necessitate the qualia themselves, thereby guarantees (for confirmation purposes) that all precise versions of physicalism will posit just this connection between the physical and the phenomenal.[5] Finally, I will explain why this assumption is unjustified, and explore what lessons can be learned as a result.

Some preliminary matters and the canonical formulation

The evolutionary argument is an inference to the best explanation, and consequently involves the evaluation of numerous different hypotheses. Before presenting the argument, it will be important to get a feel for the various hypotheses that might explain the evidence that needs explaining.

When examining the evidence the argument considers, we are trying to decide between three competing general theories on the mind-body problem — physicalism, (dualistic) interactionism, and (dualistic) epiphenomenalism. Interactionism and epiphenomenalism, as I will understand them, are robust dualist views, which deny the metaphysical supervenience of qualia on the physical.[6] Interac-

[5] This assumption has often been left implicit by defenders of the argument, but we will see below that it is required to get the argument off the ground.

[6] A prominent example of the kind of dualism that I am assuming these views are committed to is the one defended by Chalmers (1996).

tionism and epiphenomenalism differ from one another only in their views about the causal efficacy of qualia — interactionism accepts that at least some qualia are causally efficacious with respect to the physical, while epiphenomenalism denies that any qualia are causally efficacious with respect to the physical.[7] I will understand physicalism, on the other hand, as any view which accepts that qualia metaphysically supervene on the physical.[8]

I will offer a couple of brief remarks on these positions before continuing on to the argument itself. First, it should be noted that my understandings of interactionism and epiphenomenalism focus on the efficacy of qualia, not mental states generally. This is convenient for present purposes because traditional evolutionary arguments have primarily paid attention to correlations between dangerous distal stimuli and various simple, somatic experiences (as the James passage above illustrates). They have largely ignored stimuli that cause more nuanced and complicated mental states, involving emotions like fear and anger (and whatever propositional attitudes are associated with these emotions). In any case, though, insofar as emotions have a phenomenological element, they will fall under the auspices of these definitions. Second, on certain ways of classifying mind-body theories, some views I am classifying as physicalist count as dualist or epiphenomenalist. What views are these? I am thinking of various versions of property dualism that arise from concerns about multiple realizability and from related sympathy for role functionalism.[9] I classify these views as physicalist because they share what is, for the purposes of this argument, the most important feature in common with views that are straightforwardly physicalist and

[7] A more leisurely presentation of these views (and of the evolutionary argument itself) can be found in Corabi 2011. I avoid a leisurely treatment here because of spatial constraints, and because the treatment appears elsewhere.

[8] I will not attempt here to give any sort of precise characterization of the appropriate metaphysical supervenience relation. Such discussions are notoriously complicated and largely peripheral to present concerns. For a more detailed discussion, though, see Corabi 2011. As noted earlier, I also assume throughout that all physicalist views allow for qualia to play a causal role in behavior. (See my subsequent remark for how I am treating various forms of role functionalism and non-reductive physicalism.)

[9] For examples of such views, see Yablo 1992.

non-epiphenomenalist — they affirm the metaphysical necessitation of qualia by their physical neural bases and causation of physiological events in the nervous system (and ultimately behavior) by those physical neural bases. (It will become clear later why this shared feature is important.)

Now that we have seen the various general positions we are sorting through, let us examine the argument itself. Formulations of the argument have often been fairly quick and breezy, requiring the reader to fill in a number of important details and background assumptions. It will thus be useful to formulate a version from the ground up, making explicit as much detail as will be needed for our purposes, and so we will begin by looking at a formalized version of the traditional Jamesian version of the argument. (As mentioned previously, in the next section we will see how the formulation of this argument needs to be revised in light of problems unrelated to the key assumption — about the relationship between metaphysical necessitation and confirmation — but which are still of central concern.)

As noted above, the traditional evolutionary argument is essentially an abductive argument in favor of both physicalism and interactionism, and against epiphenomenalism. It attempts to show that physicalism and interactionism, because they allow for qualia to play a causal role in the physical world (including in behavior, presumably), lead us to expect the evidence we actually find, while epiphenomenalism does not. Hence, they are each confirmed and epiphenomenalism alone disconfirmed.

What is this evidence? It is of two kinds. The first is correlations between distal stimuli and qualia, and the second is what behaviors organisms display when exposed to various kinds of stimuli (and the grounding of those behaviors in the physiology of the nervous system).

In the process of formulating our canonical version of the argument, I will make use of two important principles. First, we should always use the most determinate evidence available. So, instead, of merely using evidence like 'sharp cuts to the arm result in avoidance behavior and are mediated by unpleasant qualia', we should use evidence like 'sharp cuts to the arm of determinate type *t* result in avoidance behavior of determinate type *b* and are mediated by qualia

of determinate type *q*.'[10] In addition, data about detailed physiological transitions in the nervous system of the organism should also be included (insofar as we know what they are). Second, I will view confirmation in an explicitly Bayesian fashion. Although there are competing theories of confirmation, Bayesianism has the advantage of allowing us to set up models that make it easier to visualize the confirmation process in action. Moreover, in the context of an argument like this one, the choice of a confirmation framework is unlikely to make any substantive difference, so presupposing a Bayesian framework will not involve smuggling in any controversial assumptions.

Now, the way the argument reaches its conclusion is to maintain that P(*e*/physicalism) and P(*e*/interactionism) are similar to one another and each is significantly greater than P(*e*/epiphenomenalism), where '*e*' denotes the relevant evidence about physiological transitions and correlations between qualia and distal stimuli.[11] It is a fundamental tenet of Bayesian confirmation theory that a piece of evidence confirms a hypothesis (i.e., makes it more likely to be true) if and only if the hypothesis is more likely on the evidence than the hypothesis is on the lack of the evidence. In turn, this relationship holds if and only if the evidence is more likely on the hypothesis than on the hypothesis's negation.[12] To put it more formally: $P(h/e) > P(h)$

[10] This is essentially because using less determinate evidence can lead to counter-intuitive confirmation results. It is true, of course, that we do not always use the most determinate evidence in our everyday abductive inferences, or even our scientific abductive inferences. But it will turn out that, in every context where we rely on less than fully determinate evidence, this is because there are either great practical difficulties in obtaining the fully determinate evidence, or else it is inconvenient to use such fully determinate evidence and it seems very unlikely that fully determinate evidence would lead to a different conclusion than the less determinate evidence we do use. I discuss these issues in more detail in Corabi 2011.

[11] For the sake of simplicity, I omit consideration of background knowledge here. I also intend the probabilities in question to be understood epistemically, as what are often called 'degrees of belief'. I will not attempt to tackle complicated probability issues here, however — I think the relevant notions are clear enough intuitively for the limited purposes of this paper.

[12] There are, of course, numerous qualifications to this thesis, but none of them is relevant for present purposes.

iff P(e/h) > P(e/~h), where e is the evidence and h is the hypothesis. But that is exactly what the above conditional probabilities are implying, of course — that the evidence is much more to be expected if one of physicalism or interactionism is true.[13]

As I mentioned earlier (and as the James passage indicated), the reason for drawing this conclusion is that when we examine the evidence, we are struck by two things. First, we are struck by the appropriateness of most of the behaviors we have when confronted by dangerous (and helpful) stimuli.[14] *Prima facie*, at least, this is not surprising on any of the hypotheses; after all, we would not be here if our ancestors had responded inappropriately to burns, cuts, and insect bites. But what is more interesting is the close correlation between dangerous stimuli and experiences that feel unpleasant in some hard to describe, but nevertheless very fundamental, sense. (These experiences are unpleasant not merely in the sense that they are not pleasant, but that they are positively "nasty" in their phenomenology.)

Here is where the traditional Jamesian argument gets its bite — if physicalism or interactionism were true, this "match" between qualia and stimulus would seem to be perfectly appropriate, since according to these views qualia exert a causal influence on behavior. Thus, if we (or our ancestors) felt something other than sharp pain when we were cut on the arm by a sharp knife, we would probably treat further cuts

---

[13] The argument claims that the evidence is much more likely conditional on interactionism, for instance, than on interactionism's negation, because interactionism's negation is the disjunction of physicalism and epiphenomenalism. Although the evidence would be likely conditional on physicalism, it would not on epiphenomenalism.

[14] For simplicity's sake, I will simply focus on the case of dangerous stimuli, though most of what is said can be applied straightforwardly mutatis mutandis to the case of helpful stimuli. Incidentally, there are cases where our dispositions are not so appropriate, of course. Take, for instance, many people's standing disposition to eat fatty foods when presented with them or to avoid vigorous exercise and painful immunizations. These cases are the rare exception rather than the rule, and most likely can be explained in a variety of ways. For instance, they may be explained by the fact that our ancestors lived in a different evolutionary environment than we do, that processes other than natural selection are at work in evolution, and that long-term individual survival is not always the goal of selection pressures. I will not speculate any further here, though, on how these explanatory stories might go.

too nonchalantly, or perhaps even seek them out, since whatever qualia we were experiencing would not motivate us to avoid the stimulus with sufficient urgency. Needless to say, this would quickly remove us from the gene pool! (We need not look to fanciful hypothetical examples to make this point. Although not precisely analogous, the tragic circumstances of many sufferers of congenital insensitivity to pain illustrate the dangers of being incapable of nociception.) Thus, if one of these hypotheses were correct, it would allegedly lead us to expect exactly what we find, which is what a high conditional probability of the evidence on the hypothesis indicates.

If epiphenomenalism were true, though, things would be different. Because epiphenomenalism entails that qualia have no causal influence on behavior, we get the intuition that qualia could be varied greatly without changing behavior at all. For example, an individual could easily feel ecstatic pleasure when cut by a knife, and still behave in exactly the same way as in the actual world. Thus, there would be no special reason to think the actual stimulus-phenomenology correlations would hold if epiphenomenalism were the case, hence the reason for the lower conditional probability of the evidence on the hypothesis.

A good (albeit idealized) way to think of the confirmation process is to envision each general hypothesis (e.g., epiphenomenalism) as a disjunction of highly determinate versions of that hypothesis, each of which specifies the history of the world in maximal detail.[15] Each of

[15] A reviewer objected that general hypotheses are not disjunctions of highly specific determinate hypotheses. Consider, for instance, the theory of plate tectonics. Surely it is ludicrous to suppose that the theory of plate tectonics is composed of a disjunction of a myriad of ultra-determinate theories specifying slightly different microscopic paths of plate movement. Worse, it seems preposterous to suppose that such theories would specify the entire history of the world in this level of detail! I respond by conceding that there is wisdom in this suggestion. For practical purposes, we do not specify theories at this level of detail because we do not have the time, memory, or computational capacity to concern ourselves with intricacies that will make no difference to our ability to assess general hypotheses. (This is because typically there will be no differences in what rival general hypotheses predict about events that are unrelated to the main phenomena they are designed to be theories about — a highly determinate version of plate tectonic theory can predict the movement of a specific atom in outer space just as easily as a highly determinate version of a rival "seafloor spreading" theory can, and

these determinate versions of the hypothesis will start off with an intrinsic probability, and the probability of the general hypothesis will be the sum of these smaller probabilities (since each of the determinate versions is mutually exclusive and together they exhaustively characterize the general hypothesis — if the general hypothesis is true, exactly one of the determinate versions will be true). As determinate versions of different hypotheses are ruled out by evidence that comes in, the probability that accrued to them initially will be reassigned to the remaining determinate options (regardless of what general hypotheses they are determinate versions of), maintaining their ratios to one another. So, for example, if a determinate version of epiphenomenalism with probability *x* is ruled out, that *x* will be distributed to all the remaining determinate options while maintaining their relative relationships. If there is a determinate version of physicalism, for instance, with probability *y* and a determinate version of interactionism with probability 2*y*, then the version of interactionism will inherit twice as much of the *x* as the determinate version of physicalism.[16]

---

vice-versa.) But it is important to realize that we are only making a concession to convenience when we omit detail in this way. An infinitely computationally powerful Bayesian demon with infinite memory and speed would not take such shortcuts. An indication that we are merely making a concession to convenience is that, when we are alerted to a potential difference between two versions of a general hypothesis that might lead to differences of prediction or to ontological differences in what is being posited, we have no difficulty recognizing that our old theory was ambiguous between them, and hence (in a sense) a disjunction of them. When assessing theoretical issues, sometimes it is illuminating to make all of this explicit and dispense with concessions to practicality. The present investigation is such an occasion, because dispensing with these concessions allows us to concentrate carefully on the characteristic ways that the respective general theories think about the production of behavior and its relationship to qualia. (The spirit of my remark here is similar to that in the note above on the precision of our formulation of the evidence.)

[16] A couple of remarks are in order. First, I assume that each general hypothesis is a disjunction of finitely many unique determinate versions (or at least countably many). If there is an infinity of determinate hypotheses comprising each general hypothesis (particularly an uncountable infinity), then this will introduce substantial mathematical complications that are well beyond the scope of the present paper, although I do not suspect that dealing with them would alter any of the substance of the arguments I give. Second, there are niceties that need

If we apply what I have said to the specific evidence at hand, the claim on the part of traditional defenders of evolutionary arguments is that, when we take the evidence into account, a large portion of the probability previously accruing to determinate versions of epiphenomenalism shifts to determinate versions of interactionism and physicalism (with no corresponding movement in the opposite direction). This is because a much larger proportion of these determinate versions of epiphenomenalism conflict with our evidence.

So, let us sum up our formulation of the traditional Jamesian argument. We can call the evidence we are considering here 'C' — it is roughly that humans have tended to behave appropriately in the light of numerous selection pressures (and individuals continue to behave appropriately in the light of familiar selection pressures), that there is a fairly smooth correlation between stimuli that enhance reproductive fitness and pleasure, and that there is also a fairly smooth correlation between stimuli that are detrimental to reproductive fitness and pain:

(1) A hypothesis *h* is confirmed iff P(*e*/*h*) > P(*e*/~*h*).[17]
(2) A hypothesis *h* is disconfirmed iff P(*e*/~*h*) > P(*e*/*h*).
(3) P(C/physicalism) > P(C/epiphenomenalism)
(4) P(C/interactionism) > P(C/epiphenomenalism)
(5) Physicalism, interactionism, and epiphenomenalism are mutually exclusive and jointly exhaustive.[18]

to be introduced to make the sort of process I describe here fully adequate and precise. None of these are relevant for present purposes, though, and so I omit them to avoid unnecessary technicality. For a bit more discussion of some of these issues, though, and a helpful visual aid, see Corabi 2011. See also Meacham 2008 for a similar visual aid.

[17] To keep things simple here, I omit reference to background knowledge.

[18] It should be noted that I consider the general hypotheses I have labeled 'physicalism' and 'dualism' to be agnostic on the question of panpsychism, and so all of the general hypothesis under consideration here are also agnostic on that question, since all are varieties of physicalism and dualism that take no explicit stands on panpsychism. (In using the terms in this way, I am following common usage in the literature in recent decades.) I will be setting aside panpsychist versions of the respective hypotheses, however, since dealing adequately with them lies beyond the scope of the limited goals of the present paper. It might

From (3), (4), and (5):

(6) P(C/physicalism v interactionism) > P(C/~[physicalism v interactionism])

And:

(7) P(C/~epiphenomenalism) > P(C/epiphenomenalism)

So, from (1) and (6):

(8) Physicalism v interactionism is confirmed.

And from (2) and (7):

(9) Epiphenomenalism is disconfirmed.

I offer a word on the interpretive justification for this formulation, since as I alluded to above James himself does not explicitly express his reasoning in a Bayesian fashion, but we are imposing the Bayesian formalism on his argument to ensure that it has adequate precision. (1) and (2) are background Bayesian assumptions discussed previously. (5) is undeniable, and although not made explicit by James, is a belief that it is fair to assume that he held. This leaves (3) and (4). On a first glance, someone might object that James never mentions physicalism, interactionism, or epiphenomenalism. How, then, could (3) and (4) be what he intended to express? It is important to note that, in the passage quoted above, James speaks of a "set of facts which seem explicable on the supposition that consciousness has causal efficacy" — a set of facts that includes the unpleasantness people feel in the presence of burns, wounds, and starvation. Invoking evolution, he suggests that "these coincidences are due, not to any

---

be objected that this represents an inappropriate assumption under the circumstances, since James's own all-things-considered view was panpsychist. While it is true that James was a panpsychist, his reasons for embracing panpsychism had no connection to the argument we are examining, and so his endorsement of that argument can be treated on its own terms independently of issues surrounding panpsychism.

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." He ends by claiming that "if pleasures and pains have no efficacy, one does not see… why the most noxious acts, such as burning, might not give thrills of delight..." It seems clear, then, that James is asserting that natural selection would kill off in the long run any organisms that sought out such "noxious acts". But according to him, only views that maintain that consciousness has causal efficacy can "explain" why organisms would avoid noxious stimuli. What does 'explain' mean in this context? The only plausible candidate is that it means *successfully predicts* — we can see this by contrasting his assessment of this view with his assessment of the "no efficacy" view. On the no efficacy view, "one does not see" why burning might not easily be correlated with very pleasant experiences — in other words, the no efficacy view makes no prediction about what sorts of qualia we would find paired with these stimuli. But the no efficacy view is epiphenomenalism, of course, as we have defined it. There is no single general view that holds that qualia are efficacious, however — there are really two views, one dualist and the other physicalist. These are the physicalism and interactionism of premises (3) and (4). Physicalism and interactionism, according to James, successfully predict the evidence, because they posit that people who survive the natural selection process will have unpleasant experiences in response to noxious stimuli, and hence avoid those stimuli, because this is what would have motivated their ancestors to avoid those stimuli and keep the species alive. Epiphenomenalism, on the other, does not predict the evidence, because according to epiphenomenalism human behavior throughout the evolutionary process would have been the same no matter what the qualia were; hence modern humans could just as easily feel delight at being burned as excruciating pain if epiphenomenalism were true. In Bayesian terms, this is tantamount to saying that the conditional probability of the evidence given physicalism and given interactionism is higher than it is given epiphenomenalism.[19]

Addressing the substance of the argument, (1) and (2) are (at least

[19] I am grateful to an anonymous reviewer for spurring me to discuss these interpretive issues in greater depth.

in outline) uncontroversial principles, and (5) is — as I already mentioned — undeniable. As I alluded to above, in the next section I will discuss a problem with (6) — and by extension with (3) and (4), which lead to (6) — that requires us to adjust the formulation of the argument and move its conclusion away from what proponents of evolutionary arguments have generally assumed is the sensible one to draw. Once we then have a finalized version in place, we will be in a position to appreciate the relevance of the key assumption, as well as the difficulties with that assumption that ultimately doom all arguments of this ilk.

The central traditional confusion and the key assumption

Before proceeding to the central traditional confusion about the argument, a remark about a more peripheral confusion is in order and will help to focus our attention more squarely on the heart of the matter. The reader may have noticed that, when a precise version of the "evolutionary argument" is formulated, the evidence having to do specifically with evolution is at best superfluous and at worst a serious distraction. This is so for two reasons, one fairly superficial and the other deeper and more far-reaching in its implications. First, the survival of presently living persons in the face of environmental challenges (and the characteristic qualia they receive as part of those challenges) gives us plenty of evidence in the spirit of the evolutionary evidence — it is probably true, after all, that most adults would not still be around if they felt pleasure at (and tended to seek out) burns, cuts, and insect stings. (As previously mentioned, sufferers from congenital insensitivity to pain, while not exactly analogous to individuals with "inverted pain spectra", do give us reason to suppose the fate of such people would not be promising. We do not really need evolutionary evidence to convince us of the problems with seeking out detrimental stimuli.) Second, when we imagine fully determinate versions of epiphenomenalism, physicalism, and interactionism (where the histories of the world are spelled out in full detail in these hypotheses), we see immediately that any possible physical history of events *outside the brains of humans* will be captured by an epiphenomenalist hypothesis, a physicalist hypothesis, and an interactionist hypothesis, and moreover each of these three maxi-

mally specific hypotheses will have roughly equal probability going in *ceteris paribus*. (This is because all of the views agree that qualia are uninvolved in events occurring outside the brain, and the only place the views disagree with one another is over the nature and/or causal role of qualia.) Thus, any evidence pertaining to matters outside the brain will be dealt with isomorphically by each of the three general views. (When determinate hypotheses are ruled out as a result of gathering evidence about behavior or evolution, the losses will be felt in equal proportion by all of the respective general hypotheses, and so will be returned to them in equal proportion.) It is only evidence about the brain itself (and about qualia) that have any chance of confirming or disconfirming any of the general views, since these are the only places the views will find themselves in serious tension with one another. For this reason, in subsequent discussion I will minimize my presentation of evidence having to do with evolution (and behavior), focusing instead on key evidence about the brain and about qualia. (I will still discuss the external stimuli that qualia are correlated with, however, as this will make it easier to see the significance of the information about qualia we have at our disposal).[20]

Now that we have seen where to focus our attention, it turns out that there are two reasons why the argument's traditional conclusion — that the argument is strictly an argument against epiphenomenalism — is unjustified, even granting the argument's key assumption (which will be discussed later).

The first reason is that the argument relies on evidence having to do with physiological transitions within the organism (in response to stimuli and resulting in behavior), especially evidence about physiological transitions within the brain. These transitions will either strongly favor physicalism and epiphenomenalism together, or else interactionism alone.[21] This is because we will ultimately wind up discovering either that they are in keeping with how physical entities outside the brain behave or that they are not. (In other words, we will wind up discovering that the behavior of the atoms and mol-

[20] For more detailed discussion of these points about the dispensability of evolutionary evidence, see Corabi 2011.

[21] This is evidence is closely related (but not equivalent) to evidence for the thesis of the causal closure of the physical.

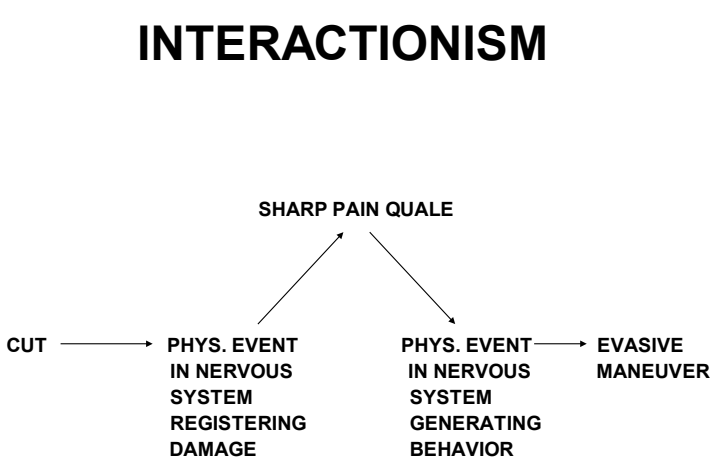
ecules inside the brain are just like the behavior of the atoms and molecules outside the brain under the same conditions, or else we will wind up discovering otherwise.) If they are in keeping with how these extra-cerebral physical entities behave, they will be the sorts of transitions physicalism and epiphenomenalism lead us to expect, since these views see behavior (and the physical events that lead to behavior) as ultimately governed by physical law alone. If they are not in keeping with the behavior of extra-cerebral physical entities, however, then this will strongly favor interactionism, since only interactionism leaves reasonable room for physical entities inside the brain to behave in a different fashion from those outside it. (This is because they are being "pushed around" by non-physical entities — namely qualia.)

The second, and for later purposes more important, reason why the argument's traditional conclusion is unjustified is that the argument relies on evidence having to do with the correlations between qualia types and distal stimuli. It turns out that when we consider the matter carefully, interactionism is subject to the same kinds of issues as epiphenomenalism where qualia "mixing and matching" is concerned — i.e., just as we have no special reason to expect unpleasant qualia to be associated with dangerous stimuli if epiphenomenalism is true, we have no special reason to expect it if interactionism is true either. This is roughly because interactionism (in most forms) posits two sets of contingent fundamental causal laws of nature where consciousness is concerned — a set of laws from physical to phenomenal (similar to epiphenomenalism), and then one from phenomenal back to physical.[22] Since they are metaphysically contingent, there appears to be no reason why these laws could not be varied to work harmoniously to produce adaptive behaviors in response to dangerous stimuli, and simply have the survival-conducive transitions causally mediated by different qualia.[23] So, for instance, if interactionism is

[22] Some forms of interactionism do posit non-mechanistic roles for qualia or other mental entities (such as, e.g., with robust agent causation views). In any case, I will set these aside for present purposes, mostly because dealing with them in full generality would take us far afield. I doubt, though, that anything about them would have a substantial impact on the basic force of my arguments.

[23] I assume throughout that fundamental laws of nature are metaphysically

true, then the actual process works like this when an organism is cut in the arm by a knife (where arrows indicate causal processes):

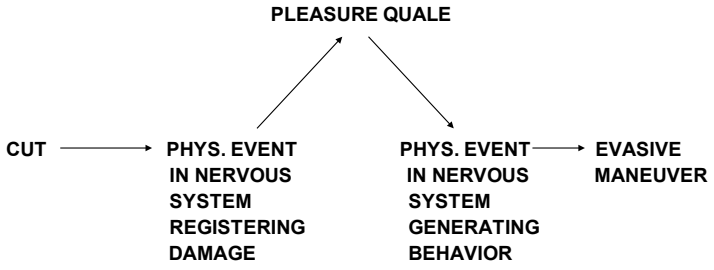


But it could have instead looked like this:

---

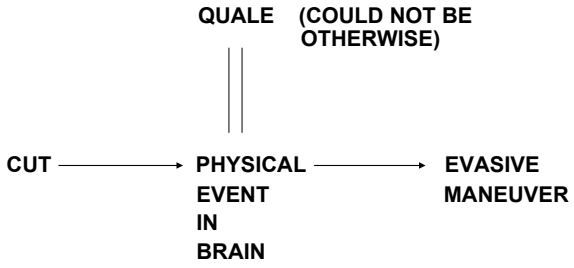
contingent, contra Shoemaker (1980), as well as what is often described as "realist" or (more informally) "oomphy" (or at least as describing oomphy causal processes). The argument can be paraphrased into a framework where the laws are treated as metaphysically necessary (so long as what properties are instantiated is not also metaphysically necessary) and perhaps also where the laws are Humean, but I will not speculate on the details of these paraphrases here. (The general idea of the necessitarian paraphrase is that there will be qualia properties that have identical "feels" to the actual ones, but which differ in their causal/nomic profiles. Thus, there will be possible worlds where such properties are instantiated, and these possible worlds will parallel the ones non-necessitarians believe in. In a standard non-necessitarian framework — where the causal/nomic profile of a property can vary from world to world — the same property would appear in many worlds, and would have many different causal/nomic profiles. In the necessitarian framework, it would be a different property in each of these possible worlds, but the centrally important feature would be preserved: the same range of causal/nomic roles matched up with the same qualitative feels.)

INTERACTIONISM ALTERNATIVE



So it is only physicalism that, by the argument's own lights, would lead us to expect the qualia/stimulus correlations we find, because only physicalism denies the metaphysical contingency between the neural base of a quale and the quale itself, as follows:

PHYSICALISM



So if the basic dialectical assumptions of the argument are correct, only physicalism will benefit from the qualia/stimulus correlation

evidence, and either physicalism and epiphenomenalism together or interactionism alone will benefit from the physiological transition evidence. But either way, it is very hard to see how physicalism and interactionism could be confirmed together and epiphenomenalism disconfirmed by itself, as the argument has typically concluded (represented in essence by (6) in the formulation of the argument in the last section). (The only way would be if physicalism benefited from the qualia/stimulus correlation evidence and interactionism benefited from the physiological transition evidence to just the right degree. But this possibility is so far-fetched, I will not even worry about it here.)[24]

So, to get a more precise feel for the relevance of these considerations, I will summarize an updated version of the Jamesian argument. Unfortunately, although many people have strong hunches about how the physiological transition evidence will turn out, at this point we have little information about the brain at a high enough resolution of detail to count as genuine evidence that can help in settling the question of whether physiological transitions will turn out to be those predicted by physicalism/epiphenomenalism or those predicted by interactionism. (In any case, as we have seen, assessing the impact of such evidence — if it does exist — is relatively straightforward.) Consequently, qualia/stimulus correlation evidence will

[24] I should briefly address an objection that may have popped up in the minds of some readers — what justifies us in supposing that if dualism is true (in either an interactionist or epiphenomenalist form), the neural base that actually generates (e.g.) a certain kind of pain could have just as easily generated a pleasure instead? I have two responses. First, there seems to be no obvious reason why not. Surely, there are some phenomenologies that would be either impossible for that neural base to generate, or at least intrinsically very unlikely — such as a complex visual phenomenology, for instance. This is because such a phenomenology would seem to require a different sort of information, or at least a great deal more information, than the neural base could reasonably encode. But why suppose pleasures and pains would be different from one another in this way? Second, and more importantly, such a variation in valence is not really required. Since it is physicalism's claim of metaphysical necessitation of the actual qualia by the actual neural bases that is doing the work, it does not ultimately matter what the character of these qualia is. All that matters is that there be a range of variations which are metaphysically possible if dualism is true, and no one would doubt that dualism allows for some variation, even if not as dramatic as flip-flops in valence.

be our main focus the rest of the way, and so I will formulate the argument so that it too focuses solely on these qualia issues. To keep things manageable, I will just suppose that the physiological transition evidence is totally up in the air, and so information about it cannot be taken into account at this stage.[25] Here is the argument, where 'Q' stands for the relevant qualia/stimulus correlation evidence:

(A) A hypothesis *h* is confirmed iff P(*e*/*h*) > P(*e*/~*h*).
(B) In general, a hypothesis *h* is disconfirmed iff P(*e*/~*h*) > P(*e*/*h*).
(C) P(Q/physicalism) > P(Q/epiphenomenalism v interactionism)
(D) Physicalism, interactionism, and epiphenomenalism are mutually exclusive and jointly exhaustive.

So, from (C) and (D):

(E) P(Q/~[epiph. v interactionism]) > P(Q/epiph. v interactionism)

And from (A), (D), and (E):

(F) Physicalism is confirmed.

And from (B) and (E):

(G) Epiphenomenalism v interactionism (i.e., dualism) is disconfirmed.

Although the above realizations damage the rationale for the standard conclusion of evolutionary arguments (i.e., that only epiphenomenalism is disconfirmed, not the disjunction of epiphenomenalism and interactionism), they leave the basic dialectical strategy essentially untouched in its core respects. Although the strategy

[25] To be perfectly satisfactory, this idea of being "totally up in the air" would have to be made more precise, but what we have should be good enough for present purposes.

does not support exactly the conclusion it was traditionally thought to, at this stage it nevertheless remains standing as a viable basis for an empirical argument designed to settle debate on the mind-body problem. Hence, from now on, I will focus on versions of the evolutionary argument that do not make the mistakes just discussed. Although these will differ from traditional versions of the argument in what mind-body theory/theories they conclude are confirmed by the actual findings (most likely, they will claim that only physicalism is confirmed), they will share with traditional versions the emphasis on the possibility of qualia "mixing and matching" to drive them to their conclusions.

Predictably, then, the qualia/stimulus correlations will be the crucial evidence in our subsequent discussion (i.e., the considerations that allegedly support (C) in the above argument). It is of paramount importance for the success of the argument that, because physicalism posits a metaphysical necessitation relation from physical neural base to quale and the alternatives do not, physicalism gains a decisive confirmation advantage where the qualia/stimulus evidence is concerned. Unfortunately, I will ultimately conclude that this crucial assumption cannot be successfully defended, and so the argument falls with it. Let us now turn to a more detailed examination of the assumption, and its bearing on the evolutionary argument.

The problem with the key assumption

The central point to note is that there is no issue about the conceptual or epistemic separability of qualia and physical events, 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 quale be associated with some other quale or no quale at all.[26] (To put things another way: we can imagine having *discovered* that the actual neural base of a certain kind of sharp

[26] I assume here that physicalism is a priori possible. If physicalism is demonstrably false a priori (as proponents of Knowledge, Zombie, and Structural Arguments have contended), then these evolutionary arguments will be unsuccessful anyway. (There may be lessons in the offing even for those who are persuaded of the truth of dualism a priori, but I will not speculate here. A bit of what I say in the conclusion addresses this issue.)

pain was actually associated with a pleasure, or with only dreamless sleep. The fact that we can imagine things having turned out this way indicates that the scenario in question is epistemically possible.) The only issue 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. Thus far, we have been following the argument above in supposing that this is not so — that metaphysical necessity is also necessity for confirmation purposes.

Reflecting a bit more on the situation, though, there does not seem to be any particularly good reason to doubt that epistemic contingency rather than metaphysical contingency should be the relevant modality where confirmation is concerned — after all, confirmation is an epistemic matter *par excellence*. Since it seems that alternate determinate physicalist hypotheses are nevertheless epistemic possibilities, ruling them out should have an adverse effect *ceteris paribus* on the likelihood of physicalism being true. This has the implication, though, that wherever some determinate version of epiphenomenalism or interactionism posits a correlation between a quale and an underlying physical brain event, there will be a parallel determinate version of physicalism that posits the same connection. This will ruin our justification for (C) in the argument of the previous section, because the only reason we had for thinking P(Q/physicalism) was greater than P(Q/epiphenomenalism v interactionism) in the first place was that these dualist hypotheses allowed for a different quale to be associated with the same underlying physical brain state (and ultimately the same external stimulus and behavior), while physicalism allowed for only the actual quale to be associated with it. This meant that, when the real quale was observed and its association with that physical brain state noted, many previous determinate epiphenomenalist and interactionist options were ruled out, but no physicalist ones were. But now, given that we recognize physicalist options corresponding to these epiphenomenalist and interactionist ones, there is a parallel process across the board, and no general hypothesis gains or loses any ground.

Objections

Before summing up the findings of the paper and discussing broader

lessons, I will deal with some objections and big picture challenges:

(1) *In spite of what you say, metaphysical possibility is really what is relevant to confirmation, not epistemic possibility.*

Response — The best way to answer this objection is to point out that it would have terribly counterintuitive consequences. To see this, consider how this approach would work in a field far-removed from the mind-body problem:

Everyone believes the identity claim 'water = H_2O' 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 the process previously discussed. But if the proposal on the table is correct, then this cannot be the right 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_2O'. 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_2O' 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_2O' had no intrinsic probability, and only can be said to have a probability at all 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_2O' had no probability along the way, then there was no probability to raise.) The ridiculousness of this conclusion is too much to stomach.

(2) It seems like the overarching complaint behind the evolutionary argument is simply the all-too-familiar explanatory gap, because all that is ultimately at issue is the relationship between physical brain states and qualia. So then why is it even worth talking about?[27]

Response — It is true that issues surrounding the epistemic separability of qualia and physical brain states wind up being of crucial importance to the evolutionary argument.[28] (This is because the argument ultimately relies on there being a crucial disanalogy between physicalism and dualism — namely, that dualism leaves it metaphysically open what qualia will be instantiated when a particular physical brain profile is instantiated, while physicalism does not.) However, there are numerous reasons the argument is worth discussing in spite of this fact. First, historically no one seems to have noticed the crucial role of explanatory gap considerations in it. Seeing that the argument has been influential (defended, in one form or another, by several luminaries), it seems worth the trouble of clarifying its relationship to other issues that are relevant to the mind-body problem. Second, although it may exploit explanatory gap considerations, those considerations are used in a very different way in the evolutionary argument than they typically are. Normally, the epistemic separability of qualia from the physical is used as the basis for a pro-dualist argument, whereas with the evolutionary argument it is used as the basis for an argument for the causal efficacy of qualia (whether those qualia are ultimately construable physicalistically or not), a largely separate matter. We even saw that once other con-

[27] I am grateful to an anonymous reviewer for spurring me to clarify the discussion of this objection.

[28] For this reason, my discussion is not meant to apply to those who claim that we can infer the presence of the relevant conscious states a priori from physical descriptions. Such physicalists are rare nowadays and I believe their position is implausible, although I readily admit that it is difficult to give convincing arguments against it (largely owing to the fact that crucial premises in any such argument would be less secure than the conclusion itself).

fusions have been unmasked and peripheral issues set aside, in this context it really forms the basis of an anti-dualist argument (albeit an ultimately unsuccessful one)!

(3) Suppose that we will eventually discover that the physical brain state underlying a particular kind of negative qualia (QN) is physical neural base 1 (NB1), and the physical brain state underlying a particular kind of positive qualia (QP) is physical neural base 2 (NB2). The negative qualia are produced by a damaging stimulus (SD) and the positive qualia by a beneficial stimulus (SB). Consider what this will do to the confirmation of the various general hypotheses. Physicalism has only two possible determinate versions to start with — version A has (SD, QN, NB1) and (SB, QP, NB2), while version B has (SD, QN, NB2) and (SB, QP, NB1). (The difference is that version B has swapped qualia valences from version A.) Epiphenomenalism, on the other hand, has four possible determinate versions to start with — the parallels of A and B and also C (SD, QP, NB1) along with (SB, QN, NB2) and D (SD, QP, NB2) along with (SB, QN, NB1). (Epiphenomenalism also allows for the swapping of what physical brain states are correlated with what external stimuli, which is what gives us the two additional options.) But then physicalism and epiphenomenalism are not parallel after all — since the probability associated with physicalism is only split 2 ways initially and the probability associated with epiphenomenalism split 4 ways, physicalism receives confirmation and epiphenomenalism disconfirmation after all once we get the final evidence, since we rule out more determinate versions of epiphenomenalism than we do physicalism.[29]

Response — A first point to make about this objection is that its presentation of the evidence (and the determinate hypotheses on the table) is incomplete; for completeness (even setting aside physiological transition evidence), we would need not just the general positive/negative valence of the qualia, but much more detailed information about their nature (and the same goes for the stimuli).

However, we do not need to dwell on these issues to see the difficulty for this objection. The main problem is that there is no reason to believe in the sort of asymmetry the objection presupposes. Why, after all, would epiphenomenalism allow *a priori* for versions that allowed for mismatches between stimulus and physical

[29] Thanks to an anonymous reviewer for suggesting a discussion of this objection.

brain state (mismatches in the sense that these physical brain states would not lead to behavior conducive to reproductive fitness) that were not allowed by physicalism? Since the neural processing of data from stimuli and the subsequent behavior generated in response to those stimuli are both a matter of the activity of physical entities governed purely by physical laws according to both epiphenomenalism and physicalism, any possible physical arrangement of the brain and nervous system in response to stimuli will be represented by a possible determinate version of epiphenomenalism and a possible determinate version of physicalism.

The upshot — evolutionary arguments fail

As I have alluded to throughout the second half of the paper, the conclusion about the space of confirmation being the space of epistemic possibility is of crucial importance. It spells doom for the evolutionary argument. Although the inferences may already be clear, it is worth spelling them out explicitly.

Essentially, what the result we have arrived at does is strip physicalism of its ability to take advantage of the metaphysical contingency of the correlation between qualia and physical events 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" by all the general hypotheses, and thus they will all remain equal to where they were beforehand. Granted, 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. To directly relate these considerations back to our updated Jamesian argument, its crucial premise — i.e., (C), that P(Q/physicalism) > P(Q/epiphenomenalism v interactionism) — is false. We have no reason to believe the evidence is more likely given physicalism than

it is given dualism.

Conclusion

At this point, we can stop and appreciate the positive lessons that can be salvaged from the demise of the evolutionary argument. Appreciating the flaws of the argument can help us to clarify exactly what considerations are potentially fruitful in helping us to solve the mind-body problem and gain a better understanding of mental causation. While perhaps not essential, purifying the discussion in this manner can help to prevent confusion and distraction in future debates.

It appears that the only directly useful empirical considerations will be ones having to do with whether physical entities inside the brain consistently behave in the same ways as those outside the brain. If they do not behave in the same ways, then for the reasons outlined above, we will have considerable evidence in favor of interactionism. And alternatively, if they do, then we will have considerable evidence against interactionism, and hence in favor of the disjunction of physicalism and epiphenomenalism.[30]

The only other tools at our disposal for dealing directly with the mind-body problem are bread and butter *a priori* considerations.[31] Surely if physicalism is ruled out or made less palatable *a priori*, this will have significant effects on the intrinsic probabilities of the determinate physicalist options, and also on the intrinsic probabilities of

[30] There is another type of evidence that could potentially play a role. If it were found that distinctive (and fairly natural, joint-carving) qualia types did not correlate smoothly with any neural base types, 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 the possibility further. (Note that the correlation in question here need only be one directional — 'if neural base n, then qualia q'. The converse sort of correlation could be ruined by multiple realizability, but this would not have an impact on the issues at hand.)

[31] 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, and information about the broad nature of the physical.

the determinate dualistic hypotheses, since probability in this setting is a zero sum game (because the general theories are together mutually exclusive and jointly exhaustive).

In any case, all other matters aside, it is clear that empirical considerations of the sort adduced by James and other traditional proponents of evolutionary arguments against epiphenomenalism will not bear fruit. Those philosophers hopeful that the empirical evidence adduced in those arguments would shed light on these issues in the philosophy of mind will either have to go back to the drawing board, or return to old fashioned armchair philosophical theorizing.[32]

Joseph Corabi
Saint Joseph's University
Department of Philosophy
5600 City Ave.
Philadelphia. PA 19131 USA
jcorabi@sju.edu

References

Broad, C.D. 1925. *The Mind and its Place in Nature*. London: Routledge and Kegan Paul.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Corabi, Joseph. 2008. Pleasure's Role in Evolution: A Response to Robinson. *Journal of Consciousness Studies* 15: 78-86.

Corabi, Joseph. 2011. Why the Evolutionary Argument isn't really an Evolutionary Argument after all. *Journal of Consciousness Studies* 18: 44-65.

Eccles, J. and Popper, K. 1977. *The Self and its Brain: An Argument for Interactionism*. Berlin: Springer-Verlag.

Howson, C. and Urbach, P. 1996. *Scientific Reasoning: The Bayesian Approach*, *2nd edition.* Chicago: Open Court.

Huxley, Thomas. 1874. On the Hypothesis that Animals are Automata. Reprinted in *Collected Essays*. London, 1893-94.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-136.

James, William. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Lindahl, B. I. B. 1997. Consciousness and Biological Evolution. *The Journal of Theoretical Biology* 187: 613-629.

[32] I am grateful to Brian McLaughlin, Susan Schneider, Audre Brokes, Jamie Hebbeler, and Todd Moody for helpful discussion and comments on previous drafts.

Meacham, C. (2008) Sleeping Beauty and the Dynamics of *De Se* Belief. *Philosophical Studies* 138: 245-269.
Robinson, William. 2003. Epiphenomenalism. In *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/epiphenomenalism/#Natural
Robinson, William. 2004. *Understanding Phenomenal Consciousness*. Cambridge: Cambridge University Press.
Robinson, William. 2007. Evolution and Epiphenomenalism. *Journal of Consciousness Studies* 14: 27-42.
Shoemaker, Sydney. 1980. Causality and Properties. In *Time and Cause*. Edited by Peter Van Inwagen. Dordrecht: Reidel.
Spencer, Herbert. 1871. *Principles of Psychology*. New York.
Van Rooijen, Jeroen. 1987. Interactionism and Evolution: A Critique of Popper. *British Journal for the Philosophy of Science* 38: 87-92.
Yablo, Stephen. 1992. Mental Causation. *Philosophical Review* 101: 245-280.