PURIFIED BY SUPERVENIENCE: A CASE STUDY ON THE USES OF SUPERVENIENCE IN CONFIRMATION

JOSEPH CORABI
DEPARTMENT OF PHILOSOPHY
SAINT JOSEPH’S UNIVERSITY
5600 CITY AVE.
PHILADELPHIA, PA 19131 USA

E-MAIL—JCORABI@SJU.EDU
PHONE—610-660-1564
FAX—610-660-1087
PURIFIED BY SUPERVENIENCE: A CASE STUDY ON THE USES OF SUPERVENIENCE IN CONFIRMATION

ABSTRACT

This paper explores the use of supervenience in clarifying and elucidating debates about empirical arguments surrounding the mind-body problem and mental causation. In particular, it examines William James’s famous evolutionary argument against epiphenomenalism, and shows that concepts associated with supervenience—in combination with Bayesianism—can transform our understanding of how the evidence the argument adduces should be handled, because it gives us a useful framework for organizing our examination. After providing a thorough critique and updating of the traditional argument using the tools provided by supervenience, objections to the general approach are also addressed.
1. Introduction

The introduction of the general concept of “supervenience” has revolutionized discussion in many areas of philosophy, clarifying and elucidating problems that people previously struggled to formulate and address in precise ways.¹ Many of these improvements are now familiar, and include clarifications of debates about the characterization of physicalism, externalism about mental content, and material constitution. My aim in this paper will be to explicate another area—arguably an underappreciated area—where introducing supervenience is extremely useful in making philosophical progress: evaluating empirical arguments about the mind-body problem. Specifically, our focus will be on William James’s famous evolutionary argument against epiphenomenalism. By using this argument as a case study, we will see the way that supervenience can clear up difficult territory and provide new insights into issues that combine philosophical and scientific components—in this case by providing a framework that allows us to organize our examination and avoid pitfalls that are easy to fall into otherwise.

In the next section, I will offer an introduction to James’s argument, providing a formulation of the argument and discussing a number of background issues. In the section to follow, I describe the ways that supervenience—in combination with a Bayesian approach to confirmation—can be used to formulate the argument in more rigorous terms. This will put us in position to appreciate several lessons that are probably quite surprising to those used to examining the argument from a traditional pre-Bayesian, pre-supervenience vantage point. In particular, we learn that (i) the evolutionary evidence that purportedly forms the crucial foundation of the argument is irrelevant.

¹ For a general introduction to supervenience, see McLaughlin 1995 and McLaughlin and Bennett 2011.
(subject to a caveat), (ii) there are two kinds of evidence that the argument’s original formulation does not explicitly address that are relevant *prima facie*, and (iii) one of those varieties of evidence is ultimately not relevant because of confusions about the role of metaphysical considerations in confirmation—confusions laid bare by the supervenience apparatus. In the end, after we purify the argument using the tools that recent discussions on supervenience provide us, we see that it ultimately leads us in a much more complex and subtle direction than its traditional proponents have supposed. Not everyone is convinced of the overall fruits of this approach, though. After we have seen the strategy set out, we will examine some objections, including one that has been put forward recently by William Robinson.2

2. The Evolutionary Argument: A Preliminary Presentation

The evolutionary argument is designed to help us decide between three competing positions on the mind-body problem—physicalism, interactionist dualism, and epiphenomenalism. We can understand physicalism straightforwardly as the thesis that the conscious mind metaphysically supervenes on physical entities in the brain.3 (In other words, it is metaphysically impossible for there to be a change in the phenomenological properties of a thing without a change in the physical properties of the

2 See Robinson 2014.
3 Throughout, our focus will be on the phenomenological character of various mental states. Whether these states have representational properties and whether these properties are “narrow” is a question that we can sidestep. (For an introduction to issues surrounding phenomenology and representation, see Horgan and Tienson 2002.) In addition, I will assume that all physicalist positions allow phenomenological states a causal role in behavior. There are technical disputes surrounding mental causation even in situations where the phenomenological states metaphysically supervene on physical bases, but these will not affect the main thrust of the arguments we will consider, and introducing them would add substantial complexity to the presentation. (For a classic discussion, see Yablo 1992.) The same goes for consideration of unusual physicalist views where even the physical supervenience bases of phenomenological properties are cut off from causing behavior.
Interactionist dualism is the view that the conscious mind does not metaphysically supervene on physical entities in the brain (it at most nomically supervenes), but phenomenological states nevertheless play a causal role in behavior. Finally, epiphenomenalism shares with interactionism its dualism—it rejects the metaphysical supervenience of mind on brain—but rejects interactionist dualism’s belief in the causal efficacy of phenomenological properties. According to the epiphenomenalist, these phenomenological properties are causally inert.

It is no secret that epiphenomenalism strikes most people as a counterintuitive view. When students first hear about it, their initial reaction is often to ask things like «what do you mean the pain I get when a bee stings me isn’t what causes me to move my hand out of the way?!» They give, in the famous words of David Lewis, the incredulous stare. But noticing that a view is counterintuitive is not particularly good evidence that it is false—the world is full of counterintuitive views that are also true, particularly when those views are about empirical phenomena. (Virtually every well-established view in modern physics may be counterintuitive, at least to the uninitiated.)

In the late 19th and early 20th centuries, William James was one of the first people to try to produce a criticism of epiphenomenalism that didn’t ultimately boil down to its counterintuitiveness. In his *Principles of Psychology*, he gives the following 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

---

4 I am introducing supervenience-talk at this point as a convenience. It is much easier to formulate the respective positions using it. The main advantages of formulations in terms of supervenience will arise later and be largely unrelated to this issue.
of food, drink and sleep, work when exhausted, burns, wounds, inflammation, the
effects of poison, are as disagreeable as… 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.\footnote{James 1890, 143-4—emphasis in original.}

The basic gist of this argument is clear. James describes three observations that
he believes are of crucial importance: (1) humans evolved, largely by natural selection,
(2) there is a strong correlation between harmful stimuli and unpleasant experiences, and
(3) there is a strong correlation between beneficial stimuli and pleasant experiences.
Having made these observations, James notes that we can easily explain them if
«consciousness has causal efficacy»—i.e., if either physicalism or interactionist dualism
is true. According to James’s thinking, if either of these theories is true, then the causal
efficacy of phenomenological properties would have all but guaranteed the correlations in
(2) and (3) were humans exposed to the natural selection process in (1). This is because,
if (e.g.) burns felt pleasant, the pleasantness would have caused humans to seek them out.
This in turn would have led our species to extinction. However, if epiphenomenalism
were true, there would be no reason why these correlations would hold. Since
phenomenological properties play no causal role in behavior according to epiphenomenalism, we could have just as easily felt ecstatic pleasure while being burned if epiphenomenalism were true, and this would not have affected our behavior in any way. Hence, we have a good empirical argument against epiphenomenalism and in favor of the disjunction of positions that allow for phenomenological properties to play a causal role in behavior—namely physicalism and interactionist dualism.

There is much that could be added to give the argument further subtlety, but the basic reading of James here captures a chain of reasoning that has struck many as plausible. We can represent it as follows:

(1) There is a smooth correlation between the utility of stimuli (harmful or beneficial) and the hedonic valence of sensations associated with them (pleasant or unpleasant).

(2) Humans evolved, largely via natural selection.

(3) If (1) and (2), then theories that claim that pleasant or painful sensations have no effects on behavior fail to explain the smooth correlation.

From (1)-(3):

(4) Theories that claim that pleasant or painful sensations have no effects on behavior fail to explain the smooth correlation.

(5) If (1) and (2), then theories that claim that pleasant or painful sensations do have effects on behavior explain the smooth correlation.

From (1), (2), and (5):

---

6 For much more on the context and interpretation of this argument, see Corabi 2011 and 2014. For ease of exposition, I use the term ‘sensation’ as a shorthand designation for the phenomenological properties that together comprise a particular experience. (Subtleties about the metaphysical nature of experience need not concern us.)
(6) Theories that claim that pleasant or painful sensations do have effects on behavior explain the smooth correlation.

(7) If (4) and (6), then physicalism and dualistic interactionism are better theories than epiphenomenalism, all else being equal (i.e., they are more likely to be true).

So, from (4), (6), and (7):

(8) Physicalism and dualistic interactionism are better theories than epiphenomenalism, all else being equal (i.e., they are more likely to be true).7

3. Supervenience, Confirmation, and The Evolutionary Argument

3.1 The Crucial Ideas

Now that we have seen a preliminary interpretation of the evolutionary argument—a formulation that is quite faithful to at least the surface of James’s text—it is time to see how it might be improved and critiqued. (By “improved” here, I mean made more rigorous, not necessarily more attractive and not necessarily preserving all the details of the original formulation.) As I mentioned earlier, this will give us insights generally into how ideas related to supervenience can benefit our understanding of arguments where empirical and philosophical issues are intertwined.

For our purposes, there are two crucial ideas that lead us to a deeper and sharper understanding of the sort of reasoning James employs.8 Both have to do with supervenience:

---

7 This formulation is based heavily on the one in Robinson 2014.
8 For some recent discussions that offer attempts to improve, criticize, or defend the argument that are unrelated to the issues we will discuss, see Robinson 2007, as well as a number of the suggestions in Corabi 2008.
(A) The use of Bayes’s Theorem (and the ancillary principles that together with
the use of the theorem comprise Bayesianism) give us a useful and precise
way to think about the process of confirming or disconfirming empirical
hypotheses. When we use Bayes’s Theorem in the confirmation process, the
ultimate epistemic probability we have in a particular general hypothesis—at
least in the ideal—logically supervenes on our degree of belief in its fully
determinate versions. (In other words, it is logically impossible for there to
be a change in one’s degree of belief in a general hypothesis without a change
in one’s degree of belief in its determinate versions.)

(B) When distinguishing different hypotheses for confirmation purposes, it is
epistemic supervenience that matters, not metaphysical supervenience.

Let us unpack these two points and appreciate their relevance to the evolutionary
argument, beginning with (A).

3.2 Specific Hypotheses and General Hypotheses

Not everyone agrees that Bayesianism is the best normative or descriptive theory of
human inference to the best explanation or confirmation. On the descriptive side, many
psychologists have claimed that human beings often do not respect the probability

---

9 See Howson and Urbach 1996 for a helpful introduction to Bayesianism.
10 I will use the terms ‘epistemic probability’ and ‘degree of belief’ interchangeably.
11 ‘Logical possibility’ here refers to compatibility with logical and mathematical laws. So, it is logically
impossible for the two premises of a modus ponens argument to be true and the conclusion false. Similarly,
it is logically impossible (again, given the appropriate idealization) for an individual to have a degree of
belief .5 in X, .6 in Y, and .4 in “X or Y.”
12 Epistemic and metaphysical supervenience will be discussed in more detail later. The key difference is
that epistemic supervenience has to do with relationships of concepts to one another, while metaphysical
supervenience has to do with relationships of properties to one another.
calculus in their empirical reasoning.\textsuperscript{13} On the normative side, theorists have argued that humans have nothing like the memory or computational power required to implement Bayesian reasoning on a systematic basis, and consequently trying to do so is not to be recommended.\textsuperscript{14} These issues aside, there is little doubt that if humans had precise memory and unlimited computational power, some form of Bayesianism would be the way to approach inference to the best explanation. Bayesianism, in other words, is the ideal of perfection to which the practice of confirmation aspires. Even given the limitations in humans’ ability to implement it, this is probably why Bayesianism is the leading normative view among confirmation theorists today and the one that receives the most scholarly attention. Thus, insofar as we can understand how a perfect Bayesian computer would reason with our evidence, we will come to insights about what that evidence supports, even if it would be a mistake in practical circumstances to try to mimic the computer to a T. This is the spirit of my use of Bayesian principles in an investigation of the evolutionary argument.

When we use Bayesian principles as part of the confirmation process, we implicitly commit ourselves to the logical supervenience of degrees of belief in general hypotheses on degrees of belief in their fully determinate specific versions, which we can think of as possible worlds.\textsuperscript{15} This is because any true general hypothesis will be made true by some maximally specific detailed history of the world. If the General Theory of Relativity is true, for instance, the General Theory of Relativity will be made true by

\begin{footnotesize}
\begin{enumerate}
\item The most prominent such scientists are probably Daniel Kahneman and Amos Tversky. See, for instance, the famous Tversky and Kahneman 1983.
\item See, for instance, Gigerenzer and Todd 1999.
\item The issue of what kind of possible world it is will be addressed below—some additional technicality is required for a perfectly precise characterization, but none of this will be relevant for our purposes. Again, we are also assuming here that we manifest the Bayesian ideal of very precise memory and vast computational power, otherwise we would not be able to attend to such fine-grained world descriptions.
\end{enumerate}
\end{footnotesize}
some very specific world history where, intuitively, the laws of the General Theory of Relativity are operative. The same goes for epiphenomenalism. There are many different maximally specific descriptions of the world that are all compatible with the truth of the general epiphenomenalist view—there are maximally specific descriptions of the world where the city of London never existed, for example, or where one atom is currently out of place (but all else is correct). Epiphenomenalism as a general view can be thought of as the disjunction of all of these maximally specific views that share in common the core epiphenomenalist claims—namely that phenomenological properties do not metaphysically supervene on physical properties and phenomenological properties are causally inert with respect to behavior. We can, of course, say parallel things for the other competing general mind-body theories. An idealized Bayesian agent will be keeping track of which general theories are most likely, but also what specific versions of those general theories are most likely as well.

These reflections lead us to a natural metaphor for thinking about the confirmation process.\textsuperscript{16} We can think of all of the possible worlds (or at least all the relevant possible worlds)\textsuperscript{17} arranged two-dimensionally on a pane of glass. The epiphenomenalist worlds are all grouped together in one region, the interactionist dualist worlds together in another, and the physicalist ones in yet another. Initially, above each tiny region corresponding to a respective world is a volume of water, and all of these

\textsuperscript{16} This metaphor is very similar to ones that appear in Meacham 2008, Bostrom 2014, 10, and Corabi 2011 and 2014.

\textsuperscript{17} To keep our presentation here manageable, I am setting aside possible worlds where all of the three main general theories are false. I am also assuming that there are only finitely many possible worlds and that there are no panpsychist worlds. Significant complexity must be introduced if these assumptions are relaxed, but none of that complexity will affect the main lessons here. Some simplifying assumptions are also being made about the confirmation process, but again none of these are germane to the issues we are exploring. For some additional detail, see Corabi 2011.
volumes of water collectively add up to 1.\textsuperscript{18} The overall sum of volumes for the worlds classified under each particular general hypothesis is the volume for that general hypothesis. (The volumes here obviously correspond to degrees of belief. This metaphor illustrates the basis for the supervenience of the degree of belief in a general hypothesis to the degree of belief in its specific versions.)

As evidence begins to come in (e.g., evidence about the history of the human species or evidence about sensory-stimulus correlations), some of the possible worlds are ruled out because they are inconsistent with the data that has accumulated. As possible worlds are ruled out, their volume shrinks to nothing. However, this water doesn’t go away—the total volume always remains 1. Rather than evaporating, the water is reapportioned to all the remaining possible worlds, preserving the ratios of their volumes relative to one another. So, for instance, suppose there are only three worlds—worlds A and B each have volume .25 and world C has volume .5. Now, imagine world A is ruled out. Its .25 goes to 0, but the .25 it had is now redistributed to B and C, keeping the 2 to 1 ratio between them the same. So, after the update, C has 2/3 of the volume (having received roughly .167 of volume from A) and B has 1/3 of the volume (having received roughly .083 of volume from A).

Notice at this point that, in order to decide whether there is an inconsistency between a world and the received evidence, we would be wise to formulate the evidence in the most specific way possible. Thus, we would not want to note merely that «most cave women have unpleasant sensations when stung by a bee». Instead, we would want

\textsuperscript{18} The initial volumes are set \textit{a priori} in some way, perhaps by simplicity. For our purposes, since we are primarily considering the impact the evidence has on the respective general hypotheses rather than on the all-things-considered likelihood of those hypotheses, we can largely ignore the vexed issue of prior degrees of belief.
to formulate evidence more along the lines of «Linda the cave woman at time $t$ has sensation with phenomenological properties $p$ in the face of bee sting of specific type $s$ and engages in avoidance behavior of type $a$». (An ideal Bayesian agent would get more specific still.)\(^{19}\) Thus, when I observe (e.g.) that I get a particular kind of pain $k$ when I cut myself in the arm in a particular way $w$ and engage in avoidance behavior of type $a$, this will be inconsistent with a variety of determinate hypotheses (i.e., possible worlds). For instance, it will be inconsistent with a possible world where that sort of cut in my arm gives me pleasure, or gives me pain of determinate sort $j \neq k$. It will also be inconsistent with possible worlds where I seek out the stimulus rather than avoiding it.

When we combine all of this, we arrive at some interesting insights. First, all of the evolutionary evidence (premise (2) in the reconstruction above and utilized conditionally in other premises) that supposedly drives the argument to its conclusion appears to be irrelevant.\(^{20}\) This is because all of the evolutionary evidence involves information about external stimuli and bodily movement. External stimuli and bodily movement happen outside the brain, and so are not associated with phenomenological properties or the areas of the body where phenomenological properties would play a direct causal role in processes affecting physical entities (i.e., in the brain).\(^{21}\) Thus all of the general hypotheses presumably have parallel specific versions that are ruled out each time we learn that a particular kind of behavior is associated with a particular kind of stimulus, because all of the general hypotheses are hypotheses about the mind and its

\(^{19}\) If we do not formulate evidence in the most determinate way possible, we can potentially wind up with clearly incorrect confirmation results. See the sick man example in White 2000, for instance.

\(^{20}\) It is irrelevant subject to a caveat that will be discussed below.

\(^{21}\) Recall that we are setting aside panpsychist possibilities. We are also setting aside views that aim to individuate phenomenal states via their associations with entities or processes outside the brain (whether elsewhere in the body or in the external environment). I assume here that there is a way to individuate phenomenal states in a principled way without appeal to these external entities.
most immediate causal impact, not hypotheses about physical phenomena that do not involve anything mental. (There will be an epiphenomenalist world, a dualistic interactionist world, and a physicalist world that each predict the pairing of this specific bodily movement with this specific stimulus. And there is no apparent reason to suppose that any will have more volume than the others. This will result in a wash once the volumes are reassigned.)

Second, we learn that it would be virtually impossible for interactionist dualism and physicalism to be confirmed (i.e., made more likely) together and epiphenomenalism to be disconfirmed alone by whatever evidence is left. To see why, consider the varieties of evidence that remain after we dispense with the evolutionary evidence. All of it has some connection to goings-on in the brain and phenomenological sphere: (i) evidence about the correlations between phenomenological properties and underlying physical brain states, (ii) information about the transitions of physical entities within the brain, and (iii) information about the correlations between external stimuli and particular phenomenological properties—e.g., bee stings of type $s$ with unpleasantness of type $p$.

For the moment, let us set aside (i)—it will be taken up in the next section, and will be exposed as irrelevant ultimately. Instead, we will focus on (ii) and (iii). With respect to (ii), both physicalism and epiphenomenalism predict that physical entities in the brain will behave in exactly the way that straightforward physical laws suggest, since according to both theories the only properties involved in causing these transitions are physical properties. Only dualistic interactionism allows for the physical entities to

---

22 Strictly speaking, the assumption is that there is no apparent reason to suppose that any will have a volume that is out of keeping with the proportion of the volumes of the general hypotheses to one another.

23 The epiphenomenalist denies any causal role for phenomenological properties (at least of a sort that could lead to behavioral changes, as these surely could). The physicalist believes that the only properties
behave other than physical law alone would predict, because only interactionism allows for non-physical entities to play a role in the causation of behavior. Hence, although we have little information at present about the nitty-gritty behavior of the physical brain, we will ultimately discover either that its activity corresponds to the wild predictions made by at least a sizeable portion of the specific versions of interactionist dualist, or (as many people conjecture) with the boring predictions made by virtually all versions of physicalism and epiphenomenalism. But no matter what, the result is almost certainly not going to line up with the standard conclusion we saw in the preliminary reconstruction above, which holds that dualistic interactionism and physicalism will be confirmed together and epiphenomenalism disconfirmed alone.

Consider (iii) now—correlations between external stimuli and phenomenological properties. (We can focus our attention on correlations in presently living individuals, even just in ourselves if we like.) Once again, it is virtually impossible that we will arrive at the standard conclusion above. The original argument gets its bite by noting that epiphenomenalism alone denies sensations a causal role in behavior, and so only it allegedly allows sensations to be “mixed and matched” while keeping the transition from stimulus to response (mediated by a variety of processes in the nervous system and—what is for our purposes salient—the brain) the same. Thus, according to this line of reasoning, only epiphenomenalism fails to explain or predict the strong correlation between harmful stimuli and unpleasant sensations, and between beneficial stimuli and are ultimately physical, and so clearly these are doing the causing in the brain—recall that we are setting aside technical metaphysical worries about mental causation for theories that embrace the metaphysical supervenience of phenomenological entities on physical ones.

24 The official claim here would have to invoke the share of volume of the respective general hypotheses associated with specific versions that made predictions in keeping with physical laws vs. ones that didn’t (rather than the sheer quantity of those specific versions). But the point should be clear enough for our purposes.
pleasant sensations. However, things are not so simple. It is true that epiphenomenalism does allow this sort of mixing and matching while leaving transitions in the brain unchanged—this is because the particular physical brain states associated with the characteristic transition from stimulus to response could give rise to any number of different sensations.\(^{25}\) Thus, there are many epiphenomenalist possible worlds with very different qualia-stimulus correlations than what we actually observe, which then get ruled out as new evidence about these correlations arrives. But epiphenomenalism is not the only general hypothesis like this. Dualistic interactionism also allows for the mixing and matching of sensations, since dualistic interactionism also posits contingent laws between physical states in the brain and sensations. In the case of dualistic interactionism, there are two sets of laws rather than one—a set of laws involved in the causation of sensations by physical brain states (this is akin to what we find with epiphenomenalism, since it is also a dualist view) and a set of laws involved in the causation of brain state changes by sensations.\(^{26}\) Together, though, these contingent laws can co-vary to produce exactly the kinds of intuitive mismatches that epiphenomenalism allows for. For instance, there is a dualistic interactionist possible world where a particular kind of bee sting produces brain state \(\beta\) in me. This in turn produces a particular kind of sharp pain (thus leading to a bee sting-sharp pain correlation), which in turn (via another contingent law) causes the brain to transition into state \(\gamma\), which ultimately leads to avoidance behavior. But there is also a

---

\(^{25}\) I am assuming here that the laws of nature are contingent, contra philosophers like Shoemaker. (See Shoemaker 1980.) If the laws are contingent, then it is metaphysically possible for the physical basis of a given sensation to cause any number of different sensations. Incidentally, the argument could be paraphrased into a framework where the laws are necessary, so long as what properties are instantiated is not also necessary. Such a paraphrase would take us too far afield here, though.

\(^{26}\) Some versions of dualistic interactionism may allow for non-lawful causation of physical changes in the brain by mental entities, particularly when the agent is involved in a process of deliberation. But since we are focused here on processes that don’t typically involve decision making (e.g., removing one’s hand when a bee stings), we can ignore these issues.
dualistic interactionist possible world where that particular kind of bee sting produces brain state $\beta$, but where brain state $\beta$ causes a type of mild pleasure instead of this type of pain. Then, there is another law that results in this type of mild pleasure causing the brain to nevertheless transition to $\gamma$, leading to the same behavior as before. Thus, even if interactionism is true, it can produce the same intuitive mismatches between stimulus and sensations that are supposedly the problematic hallmark of epiphenomenalism.

Moreover, there is no reason to think that such deviant interactionist possible worlds are rarer than similarly deviant epiphenomenalist worlds.\(^{27}\) Thus, when the evidence comes in, there is no reason to think that disproportionately more of the epiphenomenalist volume will be lost vis-à-vis the interactionist volume, and so no reason to think that epiphenomenalism will suffer from more disconfirmation.

When we consider physicalism, though, the picture changes—at least on an initial take. Because physicalism posits metaphysical supervenience of sensations on physical brain states, there is no similar opportunity for mixing and matching in the physicalist worlds. In order for the stimulus to produce the brain transitions that lead to the observed sensations, there is no other possible option but the one that occurs.\(^{28}\) To use the previous example, when the stimulus leads to brain transitions that culminate in brain state $\beta$, brain state $\beta$ is guaranteed to be associated with the specific kind of sharp pain actually

\(^{27}\) There may be subtle differences in the outcome for the two hypotheses, since there is a complicated interplay of laws in the interactionist worlds lacking in the epiphenomenalist ones. There is no apparent reason to think that this interplay will result in the standard conclusion, however. Also, again the official assumptions here are really about the volumes associated with the ruled out possible worlds, not literally about the quantity of those worlds. But expressing things in this way introduces needless complexity into the presentation.

\(^{28}\) There is, of course, the possibility of the physical laws themselves varying. But this would presumably require other major changes to the worlds. It would also be a type of variation that could be duplicated with epiphenomenalism and dualistic interactionism, since those views also believe in physical laws. Thus, even if we considered this sort of variation, there would still be a further important kind of variation that could occur if any form of dualism were true. Consequently, I set the possibility aside.
observed, because this sharp pain metaphysically supervenes on $\beta$. ($\beta$ will also lead to $\gamma$.) Consequently, it appears that, as evidence comes in, while a variety of epiphenomenalist and interactionist worlds are being ruled out, no corresponding physicalist worlds will be getting ruled out, since there are no physicalist possibilities where (e.g.) bee stings are correlated with pleasure. Thus, epiphenomenalist and interactionist worlds will be losing their volume without corresponding losses on the physicalist side, meaning that physicalism as a general hypothesis will be inheriting more and more of the volume of the other theories. This situation can be appreciated intuitively via the following diagrams. They represent the differences between how interactionism and physicalism think about the space of possibility as I have described it.\footnote{Very similar diagrams appear in Corabi 2014.}

**INTERACTIONISM**

```
INTERACTIONISM

SHARP PAIN SENSATION

STING ── PHYS. STATES ── PHYS. STATES ── EVASIVE
IN NERVOUS SYSTEM IN NERVOUS MANEUVER
REGISTERING DAMAGE SYSTEM
GENERATING BEHAVIOR

29 Very similar diagrams appear in Corabi 2014.
```
INTERACTIONISM

ALTERNATIVE

PLEASURE SENSATION

STING → PHYS. STATES IN NERVOUS SYSTEM REGISTERING DAMAGE

PHYS. STATES IN NERVOUS SYSTEM REGISTERING DAMAGE → EVASIVE MANEUVER

PHYSICALISM

SHARP PAIN SENSATION (COULD NOT BE OTHERWISE)

STING → PHYSICAL STATE OF BRAIN → EVASIVE MANEUVER
As we will see in the next section, this way of thinking about physicalism is not quite right. But as things stand now, we see much more complexity in the ultimate story than the traditional interpretation of the evolutionary argument would suggest. The evolutionary evidence basically turns out to be irrelevant, the brain transition evidence will support either epiphenomenalism and physicalism together or else dualistic interactionism alone, and the sensation-stimulus correlation evidence supports only physicalism if it supports anything at all. This result is a far cry from what has been standardly supposed, and carefully considering supervenience in a Bayesian framework is what led us to the insights.

3.3 Epistemic Supervenience vs. Metaphysical Supervenience

Our work is not finished, however. Nor are the lessons that supervenience can teach us about the evolutionary argument. Recall (B) above from 3.1—the claim that, when we distinguish different hypotheses for confirmation purposes, it is epistemic supervenience that matters, not metaphysical supervenience.

In the previous section, it appeared that physicalism had an advantage over both of the dualist hypotheses in dealing with the stimulus-sensation correlation evidence, because physicalism involves a much tighter link between underlying brain state and

---

30 One reason this way of thinking about physicalism is not quite right is that it seems to imply that, if physicalism is true, mental properties are either identical to physical properties or epiphenomenal. But recall that we are setting aside technical issues about mental causation and physicalism, simply assuming for convenience that mental properties are causally efficacious if physicalism is true (regardless of the precise metaphysical relationship between the mental properties and their underlying physical bases). This is of course not to belittle the very legitimate metaphysical issues that arise regarding the relationship between mental properties and physical ones on a physicalist picture, just to highlight that they are orthogonal to the concerns we are addressing here. (I am grateful to an anonymous referee for spurring me to explain this matter at greater length than I do in the text.)

31 See Corabi 2014 for more detail on a number of the issues above. The point in our discussion here has not been to dot every ‘i’ and cross every ‘t’, but to show as clearly as possible the ways that supervenience is useful in arriving at fresh insights about arguments such as this one.
sensation. Because both epiphenomenalism and dualistic interactionism posit
metaphysical contingency in the relationship between underlying brain states and
sensations, many specific versions of these hypotheses are ruled out when the real
relationships are observed. An analogous process does not happen to physicalism. (To
use our metaphor from above, many epiphenomenalist and interactionist worlds lose their
volume when this evidence comes in. Physicalist worlds do not lose their volume.)

To arrive at this result, however, we must make a crucial assumption: we must
suppose that degrees of belief in the general hypotheses supervene on degrees of belief in
the relevant *metaphysically* possible worlds. This is where physicalism enjoys its crucial
advantage over the alternatives (supposing that there are metaphysically possible versions
of all the various general hypotheses). However, there is strong reason to believe that
this is not correct. There are, after all, many situations where we are faced with
competing hypotheses that are metaphysically necessary if true at all. We may not know
which is true (and so metaphysically possible), but we know it is wrong to think of there
being genuinely metaphysically possible worlds that are determinate embodiments of all
the different hypotheses. Consider, for instance, scientific questions about the chemical
makeup of a substance (take water or gold). When scientists were first exploring these
questions, there were many live general hypotheses, and hence many live specific
hypotheses falling under the general categories. But not all of the live hypotheses were
genuine metaphysical possibilities—at most one was. (In our examples above, it was that
water was H₂O and gold was Au.) But clearly all of the general and specific hypotheses
were genuine options, at least initially. Hence, whatever sorts of possible worlds
scientists had degrees of belief in—degrees of belief that formed the supervenience base
of their degrees of belief in the general hypotheses—these were not metaphysically possible worlds. The sensible alternative, however, is to think of them as epistemically possible worlds.

Epistemically possible worlds are constructed out of concepts rather than out of metaphysical properties. The idea is that only properties that conceptually supervene (i.e., for our purposes, supervene a priori) on other properties do not vary from world to world. If a property metaphysically supervenes without conceptually supervening, it will nevertheless vary from world to world, in spite of its tight metaphysical link to its supervenience base. So, for example, even though the property of being water metaphysically supervenes on the property of being H₂O, it does not conceptually supervene on that property. So, there will be epistemically possible worlds where water is XYZ, ABC, etc. At least initially, these will have some volume associated with them, since we cannot rule them out a priori.

Why should we conceive of the relevant worlds as epistemically possible worlds rather than metaphysically possible ones? We have already seen the perils of treating them as metaphysical worlds. In addition, since confirmation and inference to the best explanation are epistemic matters par excellence, it makes a great deal of sense to populate the space of confirmation with entities that are clearly sensitive to a priori, epistemological distinctions.

Regarding the argument we are considering here, there is a clear consequence of thinking of confirmation in these terms. Recall that what gave physicalism its advantage above with respect to the stimulus-correlation evidence was the tight metaphysical

32 These epistemically possible worlds are similar to what Chalmers 2011 calls “scenarios.” Some of the details of Chalmers’ overall framework there are controversial, but all that matters for present purposes is the broad outline.
connection between sensations and their underlying physical supervenience bases. Because the dualist hypotheses—dualistic interactionism and epiphenomenalism—allowed for metaphysical contingency between those bases and the sensations, they were subject to disconfirmation by the evidence in a way that physicalism wasn’t. This is again because many of their highly determinate versions (the epiphenomenalist and interactionist worlds) lost volume when the evidence came in, but no corresponding physicalist worlds did. (E.g., many interactionist and epiphenomenalist worlds that paired bee stings of a specific type with pleasurable sensations.) But once we realize that there is no tighter conceptual connection between sensations and underlying physical states if physicalism is true than if one of the alternatives is true, we come to recognize that, for confirmation purposes, physicalism is in the same boat as the alternatives where the stimulus-correlation evidence is concerned. All of them allow for contingency in the same fundamental ways, and all of them are subject to similar confirmation or disconfirmation—which is to say that the stimulus-correlation evidence does not really help us to decide between them after all.

It is worth noting that once again we have been led to these insights by thinking carefully about supervenience. The use of supervenience concepts and language has allowed us both to clearly conceive and clearly state the relationships between properties whose relationship we might easily be tempted to confuse.

4. Objections

33 We are assuming here that physicalism is epistemically coherent, and thus a candidate for metaphysical possibility. If it is not, then this will spell a priori trouble for physicalism for reasons similar to those that many recent philosophers have advanced, most famously Jackson 1986 and Chalmers 1996. But here we are setting aside a priori arguments against physicalism and concentrating on empirical arguments that may have promise in settling disputes related to the mind-body problem and mental causation.
As I mentioned above, not everyone has been persuaded that supervenience really has delivered the philosophical goods where analysis of the evolutionary argument is concerned. In this section, we will examine two objections to the general approach:

(1) It is absurd to suggest that we should think of inference to the best explanation as a Bayesian process where evidence is formulated extremely determinately, and where our degrees of belief in general hypotheses supervene on a collection of extraordinarily specific versions of those general hypotheses. Imagine a working scientist—a paleontologist, say—gathering evidence and evaluating competing hypotheses. Imagine this paleontologist examining a fossil, trying to describe the placement of its every atom, even when these details are of no scientific relevance. Imagine the paleontologist then trying to divide up the competing hypotheses into extraordinarily specific alternative versions, checking each for consistency with the description of the evidence she has come up with and then updating overall degrees of belief accordingly. This is just a ridiculously over-intellectualized way to go about inference to the best explanation and evidential reasoning more generally. Hence, it should not be recommended.

Reply—This is a version of the “Bayesianism is impractical” objection discussed earlier. It is true that, because of limitations in time, computational power, and memory, human beings cannot perform the kinds of operations that an idealized Bayesian reasoner would need to. Consequently, it is unwise in practice to recommend that working scientists operate in such a fashion. However, that does not remove the force of the points made in the sections above. Nowhere was it argued that working scientists should adopt in their everyday practice the
framework discussed. However, the results of everyday practice—whatever that practice is—should be rejected when they are known to conflict with the results of a method like the one this paper has employed. This is because such a method is known to (at the very least) approximate perfection in evidential reasoning, albeit a kind of perfection that we humans cannot typically attain because of our own cognitive limitations and time and resource-sensitive practical interests. But once we admit that some variety of Bayesianism represents the ideal, it is a very short journey to the approach we have seen here. This is because Bayesian reasoning requires that degrees of belief be consistent with logic and the probability calculus, and the probability calculus demands that a set of specific scenarios that are mutually exclusive but the disjunction of which is coextensive with a general scenario have an overall probability equal to that of the general scenario. (This is essentially just an application of the elementary rule of special disjunction.)³⁴ But this is exactly what our volume metaphor above captures in the specific case of the hypotheses at play in the evolutionary argument.

On the evidence side, the response here is further bolstered by the fact that no competent scientist would accept a conclusion drawn on the basis of less specific evidence when he or she knew that the conclusion would not hold up were the evidence formulated in more specific terms. Imagine, for instance, that Carl the Chemist is examining three hypotheses A, B, and C. He has designed an experiment that he thinks will conclusively settle the issue—it involves mixing some chemicals and observing the reaction. If A is true, he expects nothing to

³⁴ Roughly speaking, the special disjunction rule states that, for any scenarios A and B that cannot be true together (i.e., P(A&B) = 0), P(A v B) = P(A) + P(B).
happen. If B is true, he expects an explosion with green smoke. If C is true, he expects an explosion with blue smoke. If his assistant conducts the experiment and reports back to him merely that an explosion occurred, he will correctly infer that B and C were confirmed, while A was disconfirmed. If his assistant then reported further that blue smoke was given off, we would have serious questions about Carl’s competence if he were to refuse to conclude that C was the only hypothesis confirmed by the experiment, preferring instead to stick with his original conclusion that both B and C were confirmed.

(2) The entire supervenience framework you set up is misleading—it distracts attention from what really matters where the evolutionary argument is concerned. After all, as William Robinson has pointed out, the correlations that are supposed to be doing the work in the argument are between unpleasantness (of experiences) and detrimentalness (of stimuli), and the same for pleasantness and beneficialness.35 But making the evidence highly determinate (and the specific hypotheses correspondingly determinate) obscures these general correlations. This is because highly determinate characterizations obscure whether stimuli are beneficial or detrimental and whether sensations are pleasant or unpleasant—they obscure them beneath an impenetrably dense layer of detail. It is moreover implausible that natural selection could explain (or be expected to explain) why bee stings lead to exactly the sort of determinate pain that they lead to. But there is very good reason to think natural selection could explain generally why bee

35 This and some material in the sentences to follow is a close paraphrase of the objection in Robinson 2014, 233.
stings would give rise to painful sensations. *(Namely that these would lead to avoidance behavior, so long as the sensations played a causal role in behavior.)*

Reply—This objection is really a set of two distinct objections. The first objection is about determinateness. The complaint is that, by getting too specific in one’s demands for either the formulation of evidence or hypotheses, some general insight of one of the overall hypotheses will be lost. This is not the case, however. The beauty of the framework that we have set up is that general insights will ultimately derive from the collective insights of the specific versions. If (e.g.) physicalism has some general advantage over epiphenomenalism that it gains from the evidence, this is because a greater number of specific versions of epiphenomenalism were invested (or, more carefully, a greater share of epiphenomenalism’s prior volume was invested) in specific versions that ultimately conflicted with the evidence. Thus, the correct result will be preserved no matter how fine-grained our descriptions of evidence and hypotheses get, because the general hypotheses are constructions out of specific hypotheses, and general formulations of evidence are at best convenient stand-ins for the more specific descriptions that could have been given.36

This leads us to the second objection, however, which is deeper. The concern is that, if phenomenological properties do ultimately play a causal role in behavior, there is just something more appropriate (and hence more probable, all else being equal) about unpleasantness causing avoidance behavior rather than seeking

---

36 It is of course true that in practice human beings, with their limited memory and computational power, might fail to notice some insight of a general theory because they are distracted by all the detail of the specific versions. But, as explained above, none of the results of this paper depend on recommending the use of the framework in everyday practice. The point is that it represents an ideal of perfection in inference to the best explanation that we should attend to if we wish to know what the evidence really supports.
behavior (and the same *mutatis mutandis* for pleasantness and seeking behavior).

If so, then the general hypotheses that allow for phenomenological properties to play a causal role in behavior will be confirmed by the evidence, because they predict that harmful stimuli will cause unpleasant qualia which will in turn cause avoidance behavior—these hypotheses will be the ones traditionally thought to be supported by the evolutionary argument. But the framework allegedly distracts our attention from this and illicitly smuggles in the assumption that either a pleasant or unpleasant sensation could just as well lead to (e.g.) avoidance behavior.

In response, I concede that this is a deep and important issue. If it were in fact true that unpleasantness were intrinsically more fitted to cause avoidance rather than seeking behavior, this would be of potential relevance for the evolutionary argument.37 If it could be shown that physicalist and interactionist hypotheses would be more likely to pair dangerous stimuli with physical brain states that gave rise to unpleasant sensations (which in turn led to avoidance behavior), then this would ruin the symmetry between all the general theories.38 It would result in a great number of specific versions of interactionism and physicalism predicting

---

37 It would not conclusively settle anything, however, because we would also need to establish that there was no similar “fittedness” between unpleasantness and whatever the physical neural base was that is associated with the processing of sensory information from bee stings and the generation of avoidance behavior. (i.e., we would have to establish that there was no reason why a particular physical neural state would be more likely *a priori* to give rise to something pleasant rather than unpleasant.) If we could not establish this, then we would have grounds to suspect that, in an epiphenomenalist scenario, the neural state that is caused by bee stings and which gives rise to avoidance behavior would be intrinsically more likely to produce an unpleasant sensation rather than a pleasant one. This would destroy the proposed asymmetry between epiphenomenalism and the other theories, because all of the theories would then lead us to expect the observed correlations. Epiphenomenalism would lead to the proposed correlations because the only feasible brain construction that evolution could produce would likely give rise to unpleasantness, even if the unpleasantness wasn’t doing any causing of behavior.

38 This “giving rise” could either be causal, in the case of interactionism, or because of metaphysical necessitation, in the case of physicalism.
the sorts of behaviors that have been observed. They would thus be confirmed by
the evidence and epiphenomenalism disconfirmed.39

There is no real reason to suppose that this is what we will discover, however.
Although neuroscientists know a considerable amount about pain and pleasure
and their knowledge continues to grow, none of it appears to be knowledge that
helps us to see why particular physical neural profiles are “fitted” a priori to the
sorts of sensations they give rise to. (Aliens with great philosophical abilities, but
with very different neural anatomy than ours, could not be expected to inspect our
physical brains and discern from that information alone whether the specific
sensations that occur as a result of bee stings had a characteristically positive or
negative feel.) While it is implausible that the underlying physical neural profiles
caused by bee stings could cause (e.g.) complex visual experiences, it is not
similarly implausible that they could give rise to pleasures. This is because those
brain states clearly do not contain enough information, or perhaps the right kind
of information, to generate complex visual phenomenology. But the same does
not appear to be true about pleasure—pleasure does not seem to require a neural
basis that encodes significantly more information, or information of a
significantly different kind, from pain.

But in any case, even if I am wrong about this deep issue, the important point is
that it would not represent defeat for the supervenience framework I have
developed. For the framework itself is neutral on this question—in fact, as in
many other contexts, supervenience concepts give us a convenient way to model

39 This is the caveat discussed above for the point about the evolutionary evidence being irrelevant. It is
only irrelevant supposing that there are not systematic differences between the general theories where
predicting particular kinds of behaviors are concerned.
the issues. If it turns out that there is a relevant asymmetry between
epiphenomenalism on the one hand and physicalism and interactionism on the
other, then the framework will show this by having a greater share of
epiphenomenalism’s volume tied up in worlds that predict different correlations than
physicalism and interactionism. The supervenience framework will make it easier
to arrive at these insights and model them. It will not be destroyed by them.

5. Conclusion
We have now seen how the concept of supervenience and supervenience language can
aid us in understanding tricky arguments like the Evolutionary Argument—arguments
that involve both empirical and philosophical components. By appreciating that our
degrees of belief in general hypotheses logically supervene on our degrees of belief in
very specific versions of those hypotheses (or at least should, supposing that we are
conforming to the confirmation ideal), we can model and concentrate on the
characteristic ways that the respective general hypotheses differ in their predictions, and
consequently on how the evidence impacts them. And by laying out different possible
supervenience relationships of phenomenological properties on physical neural bases, we
can isolate and focus on the ones that are ultimately relevant to confirmation. While it
would be possible in principle to arrive at the same insights without the use of
supervenience talk and supervenience concepts, these tools make it much easier to
formulate the issues and make philosophical progress. In over a century of discussion of
James’s argument and other related arguments, most of the lessons we discussed here
were entirely overlooked. That fact alone provides substantial testament to the usefulness
of supervenience in this sphere.
References


Meacham, C. 2008. «Sleeping Beauty and the Dynamics of De Se Belief».

*Philosophical Studies* 138, pp. 245-269.


