Felipe De Brigard[1]

In Defence of the Self-Stultification Objection

Abstract: *Epiphenomenalism holds that mental events are caused by physical events while not causing any physical effects whatsoever. The self-stultification objection is a venerable argument against epiphenomenalism according to which, if epiphenomenalism were true, we would not have knowledge of our own sensations. For the past three decades, W.S. Robinson has called into question the soundness of this objection, offering several arguments against it. Many of his arguments attempt to shift the burden of proof onto the opponents of epiphenomenalism, hoping to show that epiphenomenalism is no less stultifying than its contenders, such as dualism, functionalism, or identity theory. In the current paper I attempt to shift the burden of proof back to Robinson, and thus to defend the self-stultification objection, by offering two counterarguments against one of Robinson's objections to one of the key premises of the self-stultification objection.*

1. Introduction

Epiphenomenalism is the view that although mental events are caused by physical events, they do not cause any physical effects whatsoever (Robinson, 2007). The so-called *self-stultification objection* is a long-standing objection against epiphenomenalism according to which, if epiphenomenalism were true, we would not have knowledge of our

Correspondence:
Felipe De Brigard, 203A West Duke Building, Box 90743, Duke University, Durham, NC 27708-0743, USA. *Email: felipe.debrigard@duke.edu*

[1] Department of Philosophy, Center for Cognitive Neuroscience, Duke Institute for Brain Sciences, Duke University.

own minds. For the past three decades, W.S. Robinson has called into question the soundness of this objection, offering several arguments against it and in support of epiphenomenalism (e.g. Robinson, 1982; 2006; 2012). Many of his arguments attempt to shift the burden of proof onto the opponents of epiphenomenalism, hoping to show that epiphenomenalism is no less stultifying than its contenders. In fact, he has acknowledged this strategy explicitly at several points in time. In a recent review of his own work, for instance, Robinson sums up the core of his arguments against the self-stultification hypothesis thus:

> It would, therefore, seem that the burden of proof is on opponents of epiphenomenalism to show why that view should be thought to be any more self-stultifying than Cartesian dualism or, for that matter, more self-stultifying than a view that holds qualitative events to be identical with neural events. (Robinson, 2006, p. 91)

In the present paper I attempt to meet Robinson's challenge by offering two arguments in defence of the self-stultification objection. Specifically, I offer two arguments against one of his proof-shifting replies to one of the key premises of the self-stultification objection. To that end, I start by briefly restating the self-stultification objection, so as to locate Robinson's reply in the right argumentative context (Section 2). Next, in Section 3, I present Robinson's objection to one of the key premises of the self-stultification objection, as reconstructed in Section 2. Finally, in Section 4, I offer two counterarguments against Robinson's objection, hoping to show that the self-stultification objection is still a valid argument against epiphenomenalism.

2. The self-stultification objection

The goal of the self-stultification objection is to mobilize the intuition that if epiphenomenalism is true, then we would have to admit that we do not have knowledge of our own sensations (Shoemaker, 1975; Dennett, 1978). But how could anyone deny that, when I hit my toe against the chair, such a physical event *causes* a mental state of a pain sensation, which in turn *causes* my subsequent mental state of knowing that I am in pain — a knowledge that clearly may have behavioural (*viz.* physical) consequences? According to Robinson (1982, p. 524), the structure of the self-stultification objection is as follows (see also Robinson, 2006; 2010):

1) If there are (non-physical) sensations then either some of them cause neural events or none do.

2) Neural events are all physical.
3) The causes of physical events are all physical. Therefore,
4) Even if there are (non-physical) sensations, none of them cause neural events.
5) For any subject *S* and any *x*, if *S* non-inferentially knows [*de re*[2]] something about *x*, then *x* must cause some change in *S* which is causally contributory to *S*'s believing what he does about *x*.
6) The acquisition of a belief either results in or is partially constituted by a change in dispositions toward behaviour.
7) Changes in dispositions toward behaviour causally depend on the occurrence of some neural event(s). Therefore,
8) No (non-physical) sensations are causally contributory to any subject's believing anything about sensations. [From (4), (6), and (7).] Therefore, [by (8) and (5)]
9) For every *S*, *S* does not have non-inferential knowledge of anything about his (non-physical) sensations. Therefore,
10) Any dualism which includes the claim that we have non-inferential knowledge about our (non-physical) sensations is false.

And since we don't want to assert that we *only* have inferential knowledge of our own sensations, then either we reject that we have any knowledge of our sensations (which seems outrageous) or we reject epiphenomenalism — and, for that matter, any dualism for which the self-stultification objection is applicable.

3. Robinson's argument against premise (5)

Although Robinson has offered several arguments against different premises of the self-stultification objection,[3] the argument I want to discuss has premise (5) as its target. According to Robinson, premise (5) finds support in a rather straightforward and intuitive argument. In a nutshell, this argument — call it *argument for premise (5)* — runs as follows: if I am aware of my pain, this awareness — which is in itself a

[2] If the kind of knowledge of *x* Robinson is talking about here was just *de dicto* the premise may not get off the ground. A charitable reading of the premise asks us to read it as knowledge *de re*. Thanks to Charles Pelling for pointing this out.

[3] For instance, Robinson (1982) has also developed an interesting story regarding causality and counterfactuals, which allows him to deflect materialist criticisms against premise (3). My argument is independent of these considerations. In fact, as we will see shortly, I'm assuming that his view of causality is correct. (For an explanation of this account, see also Robinson, 1979.) Additionally, my argument is independent of some of his most recent developments, in which he objects to variations on the self-stultification objection having to do with the immediacy of our experience (Robinson, 2006) and the meaning of our phenomenal terms (Robinson, 2012).

mental state — must have been caused by the very pain I am aware of. Furthermore, since my awareness of the pain causes my subsequent pain-avoiding behaviour ('had not I been in pain, I would not have moved my hand'), and causation is transitive, then this very pain is causally responsible for my behaviour. But according to epiphenomenalism, mental states have no causal powers whatsoever. Therefore, epiphenomenalism is false.

Robinson attempts to block the conclusion of the argument for premise (5) by objecting to its first premise, namely that if I am aware of *x* (e.g. my pain), this awareness — which itself is a mental state — must have been caused by the very *x* (e.g. the pain) I am aware of. He does so with two moves. The first one is to separate the following two theses:

(A) If *S* knows that *x* is *F* (or that *x* exists) then *S* wouldn't believe that *x* is *F* (or that *x* exists) if it weren't true (or if it hadn't occurred).

(B) *x*'s being *F* (*x*'s existing) causes *S* to believe that *x* is *F* (*x* exists).

This move allows him to show that thesis (B) is not implied by thesis (A). According to Robinson, if all the anti-epiphenomenalist can say is that the knowledge of a certain sensation is caused by the sensation itself *because* such a knowledge would not have occurred had the sensation not occurred, then all we are committed to is the claim that one can establish the truth of '*C* caused *B*' by establishing that '*B* would not have occurred if *C* had not occurred'. But this principle is not an infallible strategy to determine causality. In situations in which *A* is sufficient for *B* and necessary for *C*, even when *C* occurs before *B*, it does not follow from *C*'s sufficiency for *B* that *C* causes *B*. It is logically possible that *A* produced *B* without producing *C*; indeed, from *B*'s point of view, *C*'s occurrence is redundant (Robinson, 1982, p. 527). Likewise, when it comes to the knowledge of our mental events or states: the fact that certain mental events are always followed by certain physical events — or maybe other mental events with behavioural effects (e.g. knowing that you are in pain) — gives us no reason for saying that the former causes the latter, 'for that regularity is compatible with the assumptions that (i) both the non-physical [mental] and the bodily events are caused by a physical event and that (ii) no non-physical [mental] event causes a bodily event' (*ibid.*, p. 528). (I come back to this point below.) However contentious this argument may be, I am happy to concede the point Robinson wants to extract

from it: namely that thesis (A) is all you need — and have — in the case of the knowledge of our own sensations.

Robinson's second move is to argue that we can have knowledge of our own sensations though it need not be caused by the sensation itself. He invites us then to consider the following diagrams:

$$Stimulus \rightarrow e_1 \rightarrow e_2 \rightarrow F \rightarrow e_3 \rightarrow e_4 \rightarrow Behaviour$$

Diagram 1.

$$Stimulus \rightarrow e_1 \rightarrow (e_2 = F) \rightarrow e_3 \rightarrow e_4 \rightarrow Behaviour$$

Diagram 2.

$$\begin{array}{c} F \\ \uparrow \\ Stimulus \rightarrow e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow e_4 \rightarrow Behaviour \end{array}$$

Diagram 3.

Diagram 1 depicts the interactionist version of a causally effective sensation (F), which is caused by a physical event (e_2), and which causes the physical event e_3 responsible *ex hypothesi* of our knowledge of such a sensation. Notice here that there is no physical connection between e_2 and e_3 — or between F and e_3, for that matter — and thus no physical explanation of the knowledge of F is available to the interactionist. This is not the case with Diagram 2, which depicts the identity theorist's version of the same process. However, according to Robinson, since the identity theorist explains the causal connection between e_2 and e_3 merely in physical terms, the only two properties he can allude to when it comes to explaining why we have knowledge of F at e_3 are: (i) that e_3 occurs if and only if there is an F sensation just before e_3, and (ii) that e_3 causes changes in the stimulated subject which are appropriate to the presence of F sensations. But given that thesis (A) doesn't entail thesis (B), and given that both knowledge-relevant properties (i) and (ii) are equally satisfied by the epiphenomenalist's Diagram 3, the identity theorist's version has no advantage over the epiphenomenalist's when it comes to explain how we know our own sensations. Or so argues Robinson.

4. Two counterarguments

In what follows I offer two brief counterarguments against Robinson's objection to the argument for premise (5). My hope is to show that his objection to the first premise of the argument for premise (5) is inadequate, and that we do not have reason to reject its conclusion. Consequently, this should show that premise (5) of the self-stultifica-

tion objection still holds true despite Robinson's attempt to undermine it, which in turn should give us reason to believe that the self-stultification objection is still a good argument against epiphenomenalism.

4.1. Epiphenomenalism is explanatorily disadvantageous

The first argument I offer aims to challenge Robinson's claim that the epiphenomenalist is at no disadvantage, relative to the identity theorist, when it comes to explaining our knowledge of our own sensations. According to Robinson, anti-epiphenomenalists argue that a certain sensation, *F*, must have caused our knowledge of it — that is, the mental state of knowing that one is having an *F*-sensation (call it *k(F)*) — because *k(F)* would not have occurred, had *F* not occurred before. Now, if all it was required to causally explain why we have *k(F)* was to say that *k(F)* would not have occurred had *F* not occurred, then I am willing to concede that Robinson might be right in saying that the explanation the identity theorist could provide would not represent any advantage over the explanation the epiphenomenalist could provide.

But a causal explanation of *k(F)* must not only account for whether or not *k(F)* could have occurred had *F* not occurred before, but also for the fact that *k(F)* is *of* or *about F*. In other words, a successful account of why we have knowledge of our own *F*-sensation needs to explain why it is that *k(F)* has as its content precisely the *F*-sensation. Evidently, there would be substantial disagreement among identity theorists (or dualists and functionalists, for that matter) regarding the best way to explain how the mental content of a mental state, such as *k(F)*, can refer to a particular sensation, such as *F*. After all, there is substantial disagreement as to how, in general, the intentional contents of our mental states relate to their intentional objects. However, what there seems to be little disagreement about is that the connection between the content of a mental state and the intentional object it refers to must be, at the very least, nomological.[4]

Unfortunately, by eliminating the direct causal connection between *F* and e_3 (see Diagram 3), Robinson's epiphenomenalist model has severed an obvious nomological connection between *F* and *k(F)*. At best, the model leaves this connection unaccounted for (Diagram 4).

[4] This much seems to be relatively uncontroversial among philosophers of mind, although not all agree that the nomological connection needs to be causal. For arguments as to why the connection between mental contents and their objects needs to be nomological, see for instance Fodor (1990 and 1998). For an overview of causal theories of mental content, see Adams and Aizawa (2010).

$$\begin{array}{c} F \quad k(F) \\ \uparrow \quad \uparrow \\ Stimulus \rightarrow e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow e_4 \rightarrow Behaviour \end{array}$$

Diagram 4.

In this regard, the identity theory model depicted in Diagram 2 seems to have an explanatory advantage — relative to the epiphenomenalist model — when it comes to explaining why *k(F)* has as its content the *F*-sensation, for there is already in place a nomological connection between *F* and *k(F)*, namely a causal one. Specifically, the identity theorist can state that one knows one is undergoing an *F*-sensation *because* there is a nomological connection between the content of *k(F)* and its object — the *F*-sensation — underwritten by the causal connection between e_2, to which *F* is identical, and e_3, to which *k(F)* is identical, as depicted in Diagram 5.

$$Stimulus \rightarrow e_1 \rightarrow (e_2 = F) \rightarrow (e_3 = k(F)) \rightarrow e_4 \rightarrow Behaviour$$

Diagram 5.

However, in the epiphenomentalist model, although there is a clear nomological connection between e_2 and e_3 — namely, a causal one — there isn't an obvious one between *F* and e_3. Perhaps the epiphenomenalist could suggest that *F* and *k(F)* are nomologically connected by some kind of non-causal nomological relation. Unfortunately, this move requires the postulation of an additional sort of nomological relation the identity theorist need not be committed to, rendering such an explanatory model simpler — and arguably more parsimonious — than the epiphenomenalist's. Another alternative would be for the epiphenomenalist to reject the need for a nomological connection between the content of the mental state *k(F)* and its object, the *F*-sensation. But this move is risky, for denying a nomological connection between mental contents and the intentional objects they refer to would surely bring out the same kinds of issues causal theories of content were set up to resolve, such as the problem of disjunction (Fodor, 1984) or systematicity (Fodor, 1990; Fodor, 1998), among many others.[5]

To conclude: despite the fact that there is no fully satisfactory causal account of the relation between mental contents and their

[5] How this particular problem would resurface for the epiphenomenalist if she was to deny the nomological connection between *F* and *k(F)* would take us far beyond the scope of this paper. However, the discussion in Adams and Aizawa (2010) may prove useful for those readers interested in learning more about the different motivations behind nomological and causal theories of mental content.

objects, it seems safe to assume that at the very least such a relation needs to be nomological. The identity theory model already has a nomological connection between a sensation and the knowledge of it, by way of the causal link between the physical states said mental states are identical to. However, the epiphenomenalist model lacks such a nomological connection, for it has severed the causal link between the sensation and the knowledge of it, as well as the causal link between the sensation and the physical state that brings about the knowledge of the sensation. And this, I submit, should be enough to shift the burden of proof back onto the epiphenomenalist, whose explanatory model would be at a disadvantage relative to the identity theorist's.[6]

4.2. Taking time into account

The second argument I want to offer hopes to show that there is another reason for us to be suspicious of Robinson's reply: the time component in the causal connection between events. Causation and time are intimately intertwined. As Robinson himself admits, the notion of causation he has in mind is cashed out in terms of a necessary connection between an event-cause *followed* by an event-effect. But, if that is so, then Diagram 3 should rather look like this:

$$\begin{array}{c} \qquad\qquad\quad F \\ \qquad\qquad\nearrow \\ \textit{Stimulus} \rightarrow e_1 \rightarrow e_2 \rightarrow e_3 \rightarrow e_4 \rightarrow \textit{Behaviour} \end{array}$$

Diagram 6.

In which case the aforementioned condition (i) wouldn't be met: *F* does not happen before e_3 but *simultaneously*. Robinson has, naturally, at least three possible replies to defend (i): first, he could claim that the causal connection between e_2 and *F* is simultaneous in time. At the end of the day, if *F* is immaterial, time need not work the same way it works in the physical world. This alternative, however, may jeopardize his causal argument for the separation of theses (A) and (B), for it is assumed all along that the cause *precedes* the effect, and unless Robinson gives us a different account of what 'precedes' may mean, we are forced to understand it in terms of some event happening at a time before the time in which some other event happens.

[6] It may be worth nothing that the epiphenomenalist model faces the same disadvantage relative to the functionalist model, as the nomological connection can easily be underwritten by a causal relation between two functional states. I did not discuss functionalism here because Robinson's argument focuses on identity theory alone.

A second alternative is to claim that *F* happens at a time t_x between the time in which e_2 occurs and the time in which e_3 occurs. But, if so, then one could arguably find a neural event $e_{2.5}$ at t_x such that the temporal axis transversal to *F* and $e_{2.5}$ would look exactly like that of *F* and e_3 in Diagram 6. Graphically,

$$\begin{array}{ccccccccc} & & & & & & F & & \\ & & & & & \nearrow & & & \\ \textit{Stimulus} & \rightarrow & e_1 & \rightarrow & e_{2.5} & \rightarrow & e_3 & \rightarrow & e_4 \rightarrow \textit{Behaviour} \end{array}$$

Diagram 7.

This, of course, would only move the problem one step back.

A third alternative is to shift all mental states one step 'to the right' (i.e. to the future) in time, so the knowledge of *F*, or *k(F)*, carried by e_3 actually happens while e_4 occurs. Now the picture would be roughly like this:

$$\begin{array}{ccccccccccc} & & & & & & F & & k(F) & & k(k(F)) \\ & & & & & \nearrow & & \nearrow & & \nearrow & \\ \textit{Stimulus} & \rightarrow & e_1 & \rightarrow & e_2 & \rightarrow & e_3 & \rightarrow & e_4 & \rightarrow & \textit{Behaviour} \end{array}$$

Diagram 8.

Although this alternative may work, it seems to have at least one metaphysically troublesome consequence: one's body can be entirely annihilated at time t_1 and yet one would be entertaining one last mental state at time t_2 — a consequence that, at least to me, seems not only counter-intuitive but also problematic for epiphenomenalism. After all, one of the main motivations behind epiphenomenalism is that it supposedly offers an advantage over dualism in that it does not require the postulation of metaphysically implausible disembodied mental states. As such, if the epiphenomenalist were to accept the counter-intuitive consequence that one could entertain one mental state without a brain (or a body), then she will be undercutting one of the major incentives for accepting epiphenomenalism. Either way, it seems clear to me that if we were to accept Robinson's epiphenomenalist model, we would have good reason to demand further explanation as to what the nature of the causal connection between e_2 and *F* is, given that we are now taking the temporal difference between cause and effect into account.[7]

[7] Another important question that surfaces when the temporal component of the causal interaction between e_2 and *F* is taken into account is whether this epiphenomenal cause should be understood as strictly sufficient for bringing about *F*, or whether it is a necessary but not a sufficient condition for bringing about *F*. I read Robinson as suggesting the former, but it may be possible that a weaker notion of causation could be at play here.

In sum: in this paper I tried to show that the venerable self-stultification objection is still a good argument against epiphenomenalism. Robinson tried to dissuade us from its conclusion by giving us an argument to the effect that the knowledge of our own sensations need not be caused by the sensations themselves. I tried to show that, on the one hand, such an argument may make explaining the connection between the phenomenal content of a sensation and the content of our knowledge of such a sensation[8] even more burdensome for the epiphenomenalist. And, on the other, that Robinson's version of epiphenomenalism may generate some doubtful metaphysical consequences when the time component of the causal relation is taken into account. In conclusion, it seems as though we still have good reason to believe that epiphenomenalism is rather suspicious.[9]

References

Adams, F. & Aizawa, K. (2010) Causal theories of mental content, in Zalta, E.N. (ed.) *Stanford Encyclopedia of Philosophy*, [Online], http://plato.stanford.edu/entries/content-causal.

Chalmers, D.J. (1996) *The Conscious Mind*, New York: Oxford University Press.

Dennett, D.C. (1978) Current issues in the philosophy of mind, *American Philosophical Quarterly*, **15**, pp. 249–261.

Fodor, J. (1984) Semantics, Wisconsin style, *Synthese*, **59**, pp. 231–250.

Fodor, J. (1990) *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, J. (1998) *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press.

Robinson, W.S. (1979) Do pains make a difference to our behavior?, *American Philosophical Quarterly*, **16**, pp. 327–334.

Robinson, W.S. (1982) Causation, sensations and knowledge, *Mind*, **XCI**, pp. 524–540.

Robinson, W.S. (2006) Knowing epiphenomena, *Journal of Consciousness Studies*, **13** (1–2), pp. 85–100.

Robinson, W.S. (2007) Epiphenomenalism, in Zalta, E.N. (ed.) *Stanford Encyclopedia of Philosophy*, [Online], http://plato.stanford.edu/entries/epiphenomenalism.

Robinson, W.S. (2010) Epiphenomenalism, *WIREs Interdisciplinary Reviews, Cognitive Science*, **1** (4), pp. 539–547.

However, if this was the case, one may want to know whether there are other conditions that are necessary for bringing about *F*, and what those condition may be. Either way, once time is taken into account, the causal story offered by the epiphenomenalist seems, at best, incomplete. I thank an anonymous reviewer for bringing this issue to my attention.

[8] Chalmers' version of epiphenomenalism (1996, pp. 150ff.) may avoid this move. It would involve accepting a totally different account of mental causation, nonetheless. Still, whether it can do it or not is beyond the scope of this paper.

[9] Thanks to Bill Lycan, Sara Bernstein, and two anonymous reviewers for helpful comments.

Robinson, W.S. (2012) Phenomenal realist physicalism implies coherency of epiphenomenalist meaning, *Journal of Consciousness Studies*, **19** (3–4), pp. 145–163.
Shoemaker, S. (1975) Functionalism and qualia, *Philosophical Studies*, **27**, pp. 291–315.

Paper received July 2012; revised March 2014.