The essay below was written for distribution to participants when I was invited to present at a workshop prior to the GAP-5 (meeting of the German Analytical Philosophy Society), in Bielefeld, September, 2003. The text is unaltered except for the updating of a few references.

EPIPHENOMENALISM, NATURALLY

William S. Robinson
Iowa State University
August 17, 2003

In a recent article, Anthony Rudd describes epiphenomenalism as "characteristically a theory of last resort – one into which people are pushed by the sense that all the alternatives are even less plausible" (Rudd, A., 2000, p. 60) I do not know how many thinkers have in fact been so pushed, nor do I have any clear sense of how many philosophers have espoused epiphenomenalism for any reason. I suspect, however, that there may be some truth in Rudd's sociological speculation.

For my own part, however, epiphenomenalism has long seemed to be a view that is supported by robust intuitions. In this paper, I will try to remove the sense that epiphenomenalism is a merely negative or defensive view, designed to avoid unwanted consequences. I will try instead to exhibit epiphenomenalism as the natural outcome of a line of thought that is firmly rooted in several evident and compelling assumptions.

I do not deny that there are objections to epiphenomenalism, and I do not deny that there are some intuitions that run counter to it. Since I want the outcome of this paper to be a persuasive case in favor of epiphenomenalism, I cannot ignore these sources of doubt. However, much of what I have to say about them is already easily available (see Robinson, 1999) and so I shall be relatively brief in responding to them. The responses I do give will emphasize the intuitive plausibility of the bases for the responses.

- I -

The first assumption we need is made intuitive by well-established science. This science will be familiar to readers of this paper, but it is important to pause to attend to some details. One good way into the relevant material begins with the fact that our actions involve muscle contractions. Not all of our actions involve motion, for sometimes keeping still is a significant action. But even keeping still requires many small contractions of our muscles, so we may take muscle contraction as essential to our actions. This is not to say that muscle contractions are themselves what we generally intentionally aim to bring about; that is rarely the case. But whenever we act, our muscles are involved.

Muscle contractions are caused by releases of chemicals, which in turn are brought about by the arrival of action potentials at the axonal ends of motor neurons. The propagation of action potentials is the propagation of a wave of activity, in which some ions move within the neuron, parallel to the direction of the axon, and in which others move orthogonally to that direction, across the cell membrane. The latter motions are not random oozings; they occur at gate-like structures that allow ionic passage only under particular conditions, typically induced by events immediately "upstream" on the axonal membrane.

Similar remarks hold of events on the dendritic and somatic surfaces of the neuronal cell. Potassium, sodium, and chlorine ions pass across the cell membrane, again through gate-like structures, the condition of which is modulated by the binding of neurotransmitter molecules arriving across synapses. These transactions affect the distribution of charge within and without the cell, and that, in turn, determines the generation of action potentials.

The point of dwelling on these details is to bring about a lively sense of the fact that, behind our actions, is a finely structured mechanism. It is nearly impossible to imagine that all this complexity evolved without contributing something essential to our actions. If one abstracts from the details, it may be possible to imagine an immaterial mind somehow getting a grip on the causes of our actions. But if one keeps them firmly in mind, the idea that the causes of actions could by-pass the mechanisms of neural transaction is simply not believable.

There is a further point that I will state in my own way, but which was essentially made by Wundt (1912). The contents of, say, two seconds of our conscious mental life can be summarized in a few sentences. Perhaps, if we reflect a bit, we could write a whole page that would describe those two seconds; perhaps novelists could manage three or four pages of accurate description of two seconds of their experience. But even on the most generous assumption, the informational content (in the Shannon and Weaver sense) of our sensory experience would seem to pale before the informational content of the millions of synaptic transactions that would be required if our two seconds of brain activity are to contribute to any action – even if the action is simply the reporting of what just happened in our conscious mental life. The question thus arises of how a relatively paltry informational content could determine a relatively rich one; and the evident answer is that it cannot do so.

The upshot of these reflections is that scientifically informed thinkers will naturally regard our actions as resulting from the activity of our brain's mechanisms. Questions about how we manage to keep our actions on track, over extended periods of time, in ways that sustain us in the face of novel opportunities and adversities, will naturally seem to be questions about how the mechanisms of our brain are arranged so as to work in concert. On present understandings of the brain, the idea that our actions could be other than the products of brain mechanisms, with inputs from sensory and bodily neurons, and with manifold reentrant connections, is extremely *counter*-intuitive, and the more so the more we learn about how neurons work. The intuitive view, for those apprized of the relevant background, is that actions flow from changes in our neural systems that result from sensory stimulation, reentrant activation, and internal stimulation from the reticular system.< 1 >

I will have more to say about this intuitive picture of our constitution, but it is time to turn to two other intuitive assumptions that complete an argument for epiphenomenalism. Before I can do that, however, it is necessary to make a distinction between two classes of mental states. I do not know of any way of making this distinction that will be acceptable to everyone; at the same time, it at least approximates a distinction that everyone makes in one way or another. I shall call it the difference between *qualitative events* and *propositional states*. Under "qualitative events" I include bodily sensations such as pains and itches; afterimages and aftertastes; and hallucinations and phosphenes. Under "propositional states", I include beliefs, doubts, desires, and wonderings. Perceptions and emotions are, in my view, complex: they involve both qualitative events and propositional states. For example, during a visit to the zoo, I might see an elephant. In a normal case, I would believe there is an elephant a few yards in front of me. This belief is one that I might also have if I passed around the side of an opaque-walled enclosure and then faced in the direction where I supposed the elephant to be. When I see an elephant, however, I would not *just* have this belief: I would also have a qualitative event, in which the grayness of the elephant is visually present.

The main purpose for which I introduce this distinction is to make the point that the intuitions that support epiphenomenalism are somewhat different in the case of qualitative events and the case of propositional states. The reason for this will appear in due course; for the moment, I shall focus on qualitative events.

Consider, then, a pink afterimage that you might have after fixating on something green and then looking at a white wall. You know that there is no pink spot on the wall. You do not believe that there is something pink in your brain – or, if you do, that is because you think there is always something pink there, not that there is something pink just now, when you are having the afterimage. Intuitively, the color pink is somehow involved in the situation, but it is not involved by being the color of any physical thing. You know how you produced the afterimage – you fixated on something green and then looked at the wall. If you know something about the physiology of color perception, you will have some ideas about the neural events that produce this occurrence of pink. But that knowledge will not remove the intuitive appeal of the idea that the color pink is occurrently manifested, but is not manifested by being the color of any physical surface.

Commitment to materialism opposes this intuitive view, and it is necessary to give a very brief review of three main lines of thought on which such opposition may rest. (1) One may follow Dennett in saying that there is, in fact, no pink present, but only a 'proto-judgment' to the effect that something pink is present. I think I will have many philosophers, including many materialists, on my side if I point out that this is about as *counter*-intuitive a view as one could advance. Others may say (2) that the color pink is, in fact, identical with the *property* of being a neural event of a certain kind. In Robinson (2004), I explain why it is impossible decisively to refute this proposal; but if we have to choose between saying pink is identical with a neural property and that it is not so identical, I think the intuitive view is clearly the one that denies such an identity. In considering this case, it is important to be clear that the claim at issue is *not* the

claim that something that exemplifies pink is the same *thing* as a thing that exemplifies an occurrent neural property. This latter proposal can be modeled in more familiar terms by considering certain musical instruments, e.g., harps, which are highly resonant, and also visually distinguished by their decoration, or by their shape. The resonant instrument is the same as the shapely instrument, but the shape *property* is not the resonance property; indeed, it hardly seems to make sense to say so. And that is how it is in our case: it hardly seems to make sense to say that *pink* is the same property as being a neural event in which, for example, ratios of firings of neurons are 5:3:7:2:13:8. The intuitive view is that no neural event property is the property pink.

Some materialists give backhanded recognition of this intuition by attempting to explain it away. The clearest attempt of this sort that I know of is Papineau's (2002) "conceptual dualism", according to which phenomenal concepts are different from material concepts (e.g., neuroscientific concepts), although there is just one property of which both are concepts. What is special about phenomenal concepts, and what is supposed to explain our intuition that they are concepts of something other than neural states, is that when phenomenal concepts are used, the property of which they are concepts is actually instantiated (at least imaginatively); whereas, this is not the case when we use material concepts.

I reject this explanation for the following reason. Suppose I use the phenomenal concept *pink*. Then, according to Papineau, the phenomenal property *pink* is instantiated – I am either introspecting an experience of the right kind, or I am imagining something pink. The problem I see for Papineau's account is that the instantiated property is supposed to be the same as some material (i.e., neural) property, but nothing neural is present in my experience. Papineau's account has no way of accommodating this fact. If there is just one property and it is present, then it is present. It seems that Papineau would have to say something like: It is present *as* pink, but not *as* neural state such-and-such. But there is nothing in his account that allows him to make any such distinction.

This explains why I am not moved by Papineau's attempt to explain away the intuition that the qualities in our qualitative events are not identical with neural properties of our brains. But the main point that is to be taken forward for the rest of this paper is that, even on Papineau's view, it really is intuitive that the pinkness in, say, an afterimage is not the same property as the property of being a certain ratio of neural firings, or any property that can be defined in terms of the temporal or spatial distribution of neural firings.

The most likely materialist response to this intuition is (3) that pink is *represented* in qualitative events. If that is so, then we do not have either to deny that pink is *somehow* involved in an afterimage or to affirm that it is identical with a property of neural events. The trouble with the representationalist response, however, is that we also represent states of affairs to ourselves in our nonperceptual beliefs. For example, I now believe that some roses are pink, but I am not seeing any pink roses (or anything at all that is pink). Nor am I having any visual experience or imagination of pink. So, while representation may indeed be present in having an afterimage, simply to leave matters there is to *underdescribe* the situation; it is to leave out something that is also present. It is extremely intuitive that, in perception or in afterimages, pink is represented by

pink, and not merely by a word, or any other kind of non-pink occurrence that is merely correlated with pink.

I am well aware that materialism in general, and representationalist forms of it in particular, have been defended at length by extremely able philosophers. There is no possibility of adequately defending the rejection of these views in any single paper, much less a part of a section of one. The central point that I want to carry forward from my brief review of the three materialist responses is that materialists are not defending common sense against an intuitive monstrosity. Quite the contrary: It is the materialists who have significant work to do, because natural intuitions lie against their side.

Let us put our intuitions together. (1) Our actions – including, of course, our linguistic actions of whatever kind – are caused by our muscular events, which are caused by neural events, which are caused by neural events, . . . which are caused by inputs from our bodies and our senses and reentrant connections. (2) Many qualities in our qualitative events are really exemplified, but are not identical with properties of neural events, and are not sufficiently accounted for by the representational properties of neural events. The conclusion that these intuitively plausible views suggest is that the qualities of our qualitative events are not causes of any of our actions, including those linguistic actions in which we report our qualitative events.

To turn this suggestion into a firm declaration, we need one more assumption, namely that our actions are not routinely overdetermined. Two brief points will, I think, suffice for our present purposes. (A) Even in its abstract form, overdetermination is a *counter*-intuitive proposal. If one accepts the first intuition (i.e., that our neural events are *sufficient* causes of all our actions) the natural conclusion will be that there is no room for anything else to make a causal contribution. One need not deny that someone could be dying of a heart attack brought on by fright at the moment a bullet also enters the heart; the problem here is to make such an analogy plausible in the cases at hand, e.g., in cases of reporting on the qualities in one's experience. (B) The point I attributed to Wundt, namely, the simplicity of our experiential contents relative to the complexity of the neural events that would be required to cause a report, also speaks against overdetermination. That is, raising the abstract possibility of overdetermination does nothing to explain how the sensation of pain or the pink in an afterimage should be able to get a grip on the myriad ionic transactions that would be necessary to cause neurons to coordinate our production of the word "pain" or the word "pink".

I conclude that, if we know the relevant science, it will be natural and intuitive to think that neural events, and only those, cause our actions, including linguistic actions. And it is in any case natural and intuitive to regard the property of painfulness or pink as distinct properties from neural event properties such as ratios of firing rates or patterns of times of arrival of action potentials at particular places in the brain. And if we put these intuitive views together, the outcome is epiphenomenalism.

- I I -

Propositional states remain on our agenda, but before turning to them I want to respond to two popular objections to epiphenomenalism. The first of these may be put this way: Epiphenomenalism cannot be a natural and intuitive view, because we have a strong and natural intuition that implies its falsehood. Namely, we have a strong and natural intuition that our qualitative events are causes of our experiential reports.

My reply to this concern is that in fact we have no such intuition; or, equivalently but less starkly, the intuition we do have in this area is often misdescribed. The intuition we do have is that we would not, in general, be making the experiential reports we make if we were not having the qualitative events that we are having. *This* intuition is fully in line with epiphenomenalism. In general, we would not be making the reports we make if we did not have the neural events that cause them, and to have those neural events is normally to have the causes of our qualitative events. It is, of course, compatible with this point that there can be occasions on which there is an abnormality that results in a false report. This possibility is open because the causal chain between the neural cause of a qualitative event and a report is more than one step long. That fact implies that there can always, in principle, be ways of producing an intermediate causal step (and thus, eventually, a report) by a means other than the usual one. The significance of this point will perhaps be clearer if we observe that it holds equally for identity theory and for interactionist dualism. Taking the latter case for illustration, suppose that mental event M causes neural event N1. N1 causes N2, which causes N3, which causes muscle contractions sufficient for uttering "I have an M experience". This scenario allows, in principle, for N2 to be produced by some means other than N1, in which case the resulting report might be false.

These remarks explain why no theory can guarantee that there will be no false experiential reports. But the same evolutionary principles that result in our perception being generally reliable also have the effect of making our experiential reports generally reliable. So, it is part of epiphenomenalist theory (as it is of other theories) that, in general, we would not be making the experiential reports we do unless we were having experiences of the kind we report.

But this fact is misdescribed if it is turned into an alleged intuition that our qualitative events *cause* our reports. One relevant point here is that while causation may imply satisfaction of counterfactuals, satisfaction of counterfactuals does not imply causation. Every philosopher knows this perfectly well. In general, cases of common cause will support counterfactuals between the common effects, but, of course, not causation. Another relevant point is that no one believes that anyone *has* to report a pain, or a pink afterimage. These events are available for report, but are not linked to reporting in a way that bypasses the rest of the cognitive system. The significance of this point is that it shows that the causal chain from neural event causes of qualitative events to reports is complex; and this point, in turn, helps us see the unlikelihood of there being anything intuitive about our appreciation of the causal situation. And that is the third point. There is nothing whatsoever that we know *intuitively* about the causes of our reports. We don't even know there are neural events *intuitively*. There is simply nothing that presents itself

intuitively about the relation between a pain or a pink afterimage and neural events that will lead to a report.

I turn to the second objection that is likely to arise. This is that epiphenomenalism is self-stultifying. The charge is that if epiphenomenalism were true, I would not know that I had qualitative events; so, although epiphenomenalism might conceivably be true, it makes no sense for a person to claim to know that it is true.

Behind this venerable objection lies the assumption that where there is knowledge of something, the thing known must have an effect on the knower. This, I concede, is a natural assumption. It is so, because it is natural to take perception as a model for knowledge in general, and it is true that if I know something by perceiving it, the thing perceived must have an effect on my sense organs.

But I deny that perception is the right model for knowledge of our qualitative events, and I deny that in every case what is known has to cause something in the knower. And, despite the naturalness of the perceptual model of knowledge, I hope to be able to show that these denials are intuitively plausible. There are two considerations I will mention.

The first point recurs to the fact that there must be several steps of neural transaction between an experience and a report of it. Even if one were to suppose that qualitative events caused some neural events, there would have to be a cascade of neural causation consequent upon those initial neural events in order for a report to result. The status of the eventual report as *knowledge* depends on the reliability of this neural sequence; that is, it depends on the assumption that the end product (the report) would not be occurring if the qualitative event that is reported had not occurred. Moreover, the reliability of this counterfactual is the *only* thing we have any reason to suppose is transmitted along the chain of neural causation. But this same reliability is ensured by the assumptions of epiphenomenalism – that is, epiphenomenalism supports the counterfacutal that, in normal conditions, the report would not be occurring if the reported event were not occurring. This background makes it intuitively plausible that epiphenomenalism meets the conditions for knowledge.

If causation by the object of knowledge is denied in *perceptual* cases, then knowledge fails, because if a perceptual report is made without causation by the object reported upon, one is just guessing, and there is no reason to accept that that particular report would not have been made if the object had not been as reported. But the denial of causation that is made by epiphenomenalism does not have any such consequence. So, at least one reason that might have made one think that causation is required for knowledge does not support that claim in the case at hand.

The second point that undercuts the perceptual model for knowledge of qualitative events is that perception admits of a distinction between what is perceived and how it

looks/sounds/feels/smells/tastes, and so on. But the latter items do not; there is no way a look looks, a sound sounds, and so on. So, if we *know* how a thing looks, etc., we should not expect a model to apply to that knowledge that is the same as the model that applies to perception.

Let us go into this point a little more deeply. A noise in the physical world is a compression wave. The brain's job is to detect such waves, and to represent them in such a way that useful similarities and differences among numerically distinct noises is preserved in their representations. These representing brain states are, of course, not themselves compression waves. Nor are the sound qualities that are exemplified in our auditory experiences. These facts illustrate a general feature of perception: In perception, targets are represented by something whose own nature is different from that of the target. Knowledge in perception is knowledge by means of reliable correlation of intrinsically different kinds of things, where the reliable correlation is maintained by causal connection.

If one applied this model to knowledge of one's qualitative events, the result could be illustrated as follows. The sound qualities exemplified in our auditory experiences are represented by something that has a different nature, which is reliably correlated with the sound qualities, where the reliable correlation is maintained by causal connection.

But there is something wrong with this picture. One way to express the difficulty is to note a suspicious duplication: There will now have to be two neural states in ordinary, conscious perception – one to represent the distal object, and one to represent the experience of it. Another way to express the difficulty is to note that the picture courts regress: If we have to have a representation of our experience, that is distinct from our representation of the distal object, why do we not also need a representation of the representation of our experience, and a representation of that, and so on?

The natural reaction to both difficulties is to cut off the duplication before it begins. But that is just to say that we should not apply the perceptual model to our knowledge of experiences. At the same time, we do want our reports of experiences to express knowledge of them. How is that to be achieved? Epiphenomenalism supplies an elegant solution. For each N in some set of neural events, {N}, it will be true that (a) N represents a property of some distal object, and (b) N causes an experience of a certain kind, and (c) given normal linguistic training, N is reliably connected to ability to report on either the distal property or the property of the experience. Whether a report will actually ensue depends on current task demands, which are determined by other inputs to our cognitive systems. Likewise, whether the report will be phrased as a report on a property of a distal object or a report of our experiences will depend on task demands imposed by other factors. The report will express knowledge in either case. And, while perceived objects cause N, the experiences on which we can reliably report do not. The reliability of the connection between experiential reports and experiences is parasitic on the reliability of the connection between perceptual reports and perceived objects; but the causal structure behind these reliabilities is different, in just the way that epiphenomenalism proposes.

- I I I -

I turn now to propositional states, and I will begin by saying why these have to be treated differently from our qualitative events. Readers will, I hope, recall how counterintuitive it is to hold that the pink in an afterimage is the very same property as a neural property, e.g., exemplifying a set of firing rate ratios. But an analogous claim for propositional states is not counterintuitive at all. This analogous claim is that dispositional propositional states are *realized* by the connectivity of brain parts, and that occurrent propositional states are realized by neural activation states. For example, Jones's having a standing belief that P is just Jones's neurons being in one of a class of connectivity states, and Jones's now inferring something from P is just Jones's brain undergoing one sequence in a certain class of sequences of activation states.

Some philosophers who reject epiphenomenalism are not satisfied with this realization story. They worry that, according to it, all the behavioral work is done by properties of neural states, and the fact that these neural states embody beliefs drops out as irrelevant.< 2 > But I think materialist philosophers should not be bothered by such worries. I will briefly explain why by considering the analogy of a pump.

Each pump will have properties of design, size, and materials that are not shared by other pumps. In every case, the fluid-moving effects of a pump can be accounted for by reference to the particular characteristics of its parts, and the effects of those parts on the particles of the fluid that gets moved. So, in accounting for the effects of a pump, we need never mention the property of being a pump, i.e., we *could* (in principle) choose to couch all our explanations in relative micro terms. And if we proceed analogously to those who worry about epiphenomenalism in the way described in the preceding paragraph, we will say that being a pump is epiphenomenal, i.e., makes no causal contribution, because all the real causal work is done by properties at a (relative) micro-level.

But I think it is absurd to say that pumps don't cause fluids to move. Alternatively, if all "epiphenomenalism" about thoughts amounted to is that the property of believing (or, desiring) P has a status analogous to that of being a pump, then the "charge" of epiphenomenalism should not be regarded as a worry, but should be enthusiastically embraced by all. In short, I think that the realization story is a good story, within the framework in which it is usually offered, and that the usual "epiphenomenalistic" worries that are expressed about it are somewhat misguided.

This stance has two consequences for the themes of this paper. The first is that materialists can say something very plausible in the case of propositional states that they cannot so plausibly say in the case of qualitative events. That is, they can plausibly say that propositional states are realized by neural states, but they cannot so plausibly say that the properties in qualitative events are realized by properties of neural states. This difference is the reason why qualitative events and propositional states have to be approached a little differently by epiphenomenalists. And the second consequence is that when I say that we should be epiphenomenalists about propositional states, I mean something quite different from the "epiphenomenalistic" views that I have just

been discussing. And, of course, the reasons for what I have in mind are also quite different; as I shall now begin to explain.

My term "propositional states" includes beliefs and desires. I think we often say something true, when we say "Jones believes that P" (with some actual sentence substituted for 'P') or "Jones wants Q". (I am letting 'Q' abbreviate the more accurate but cumbersome phrase, 'that Q become the case'.) For example, most of us believe that Ecuador is in South America, and from time to time we want it to become true that tasty food is before us with no obstacle to our eating it. So, I affirm that we have propositional states. However, I also believe that there are widespread misunderstandings about believing and desiring. These misunderstandings are characteristically expressed in terms of *thoughts*, which are generally supposed to be "occurrent" beliefs and desires. To preserve links with discussions of others, I will use this terminology, despite my belief that there is often something fishy about the views that are couched in it.< 3 >

Let us focus the discussion of thoughts and epiphenomenalism by considering a claim that is unlikely to be disputed, namely,

(WT) Our words express our thoughts.

Indeed, they must, for words by themselves are merely arbitrary noises. They wouldn't mean a thing unless they were "animated" by the thoughts "behind" them, i.e., if they were not characteristically used to express our thoughts.

I will not backtrack on this agreement with (WT), but I think that what it suggests (as opposed to what it literally says) can be highly misleading. To find our way into the difficulties, we may note that (WT) leads naturally to two questions, "What is a thought?" And "What is it for words to *express* a thought?" I shall proceed here by considering three candidate views that respond to these questions.

The first view I shall consider is motivated by an observation about cases in which I say things of the form (a) "It just occurred to me that P", or (b) "The possibility that P simply never crossed my mind". Namely, in many such cases, the form of expression in (a) is triggered by my having just *said* "P" to myself in subvocal speech; and the form in (b) occurs when I do not recollect having subvocally said "P", and believe that I would recall having subvocally said "P" if I had done so. These facts suggest that recalling (failing to recall) a thought amounts, at least in many cases, to recalling (failing to recall) a bit of subvocal speech; and that view in turn suggests the first claim of what I will call "Matching Theory", in its first version.

(MT1.1) An occurrent thought is often just the occurrence of a subvocal saying.

I say "often" because I think a natural extension of the view would include images as occurrent thoughts. For example, if I have a visual image of a pitcher of ice water on a hot day, I might well report later on that I had a desire to quench my thirst, and feel that I know that through

having a memory of that image. However, for the sake of brevity, I will henceforth discuss only the more frequent case of subvocal saying.

Now, what about "expressing"? The leading idea of what I am calling Matching Theory (in both versions) is that expression is a *causal* relation between two items that have contents that match. The following statements flesh out this idea.

(MT1.2) An overt utterance expresses a thought if and only if that thought causally contributed to the utterance, where a thought has a content, and the content of the utterance matches the content of the thought it expresses.

Philosophers of language have called our attention to many questions about what exactly it is that one is saying when one utters a sentence. For example, an ironic "No" may mean "Yes, of course". I cannot go into such complications here, so I will just assume that someone who agrees with (MT1.2) will be able to exhibit hard cases as all involving special additions to basic cases for which "content of the utterance, 'P'" seems unproblematic. For example, Jones, who has recently been shopping, says "The price of eggs has gone up". The sense of (MT1.2) is that this sentence has the content that the price of eggs has gone up, and that Jones's uttering it expresses Jones's thought if and only if Jones had a thought that had the same content and that causally contributed to Jones's utterance.

Speaking is an action, but it is not the only kind of action. Since actions, in general, can be said to express thoughts, we need a parallel of (MT1.2) that is more general.

(MT1.3) An action expresses a thought if and only if that thought causally contributes to it, where the thought has a content, and the action is appropriate in light of that content.

Here again, there are many difficulties in explicating the notion of an action's being *appropriate* in light of a content. As before, I will not go into these difficulties. I am, after all, not arguing for MT, but only explaining its fundamental idea. So, I will just assume that it can succeed in making sense of the idea of appropriateness to content. For example, I will assume that we can make some good sense of the claim that turning on an air conditioner would be an appropriate action if one were uncomfortably warm and believed the air conditioner would work if turned on.

I hope that I have said enough to indicate how (MT1.2) and (MT1.3) could appear attractive; and likewise, I hope that readers will be able to see how one might be tempted by (MT1.1). But if we put these three together, we get a view that I think most will find intuitively repugnant. For it is a consequence of these three that my overt utterances express my thoughts only if they are causally dependent on subvocal sayings with the same content; and this consequence is manifestly false. It is true that I *can* repeat out loud something I have just said to myself. But my ordinary speaking is not like that. I am not, in general, reading off what I say out loud from my subvocal speech. I do not ordinarily have subvocal speech when I am speaking out loud, and when I do, it

is usually subvocal speech that runs counter to what I am saying out loud – e.g., I notice some infelicity of expression, a need to add something later, and so on.< 4 >

So, MT in version 1 is anything but intuitive. It cannot be *that* collection of views that explicates "our words express our thoughts", if the latter is to be intuitively plausible. But there is another view that has many of the same motivations, and that is less repugnant. This second version of MT accepts parts 2 and 3 of the first version, but gives up (MT1.1). We still need thoughts with matchable contents, however, and if these are not subvocal sayings, what are they? Version 2 of MT says

(MT2.1) Occurrent thoughts are unconscious events that have a content that can be matched by the content of utterances.

(MT2.2) and (MT2.3) are the same as (MT1.2) and (MT1.3), respectively.

Although this combination of views is not so repugnant as the first version of MT, it does raise some difficulties. One is that if occurrent thoughts are unconscious events, it becomes problematic how we are to introspect them. But perhaps this can be answered. For example, perhaps a thought that P is introspected if it causally contributes to a subvocal saying of "P", or if it causally contributes to the production (subvocally or overtly) of the sentence "I think that P".< 5 >

A second problem, however, cuts more deeply. To see how, let us first note that (WT) is supposed to be crushingly obvious. But it is very odd that it should be crushingly obvious, if, in order not to be obviously false, it has to rely on the postulation of unconscious causes with content. In fact, this situation is so odd as to require explanation itself. The one that seems most likely to me is this. Postulating unconscious causes with content that matches our utterances is thought to be plausible only because it is thought that such causes are *required* in order that our utterances "make sense" – i.e., that they be non-accidental, or not merely thoughtless babbling. (Just think about the phrase "thoughtless babbling" – what does this mean, if not verbal production that is not 'guided by' (i.e., causally dependent on) an immediately prior representation of the content that the words express?) For non-linguistic actions, the corresponding view would be that in order for our actions to be "appropriate to" (the content of) our thoughts, there *must have been* an event in the action's immediate past, that is 'guiding' the action, and in which the content to which the action is appropriate was represented.

But if this is the right explanation, then the Matching Theory, even in version 2, is in serious trouble. This is because it has no way of cutting off an unfortunate regress. If prior representation of a content is *required* in order for an overt utterance to be non-accidental, not "unthinking", then prior representation of a content is required for a subvocal saying to be non-accidental, not unthinking. And the same point applies to an inner saying that is modeled on subvocal speech, even if it is held to be unconscious. If prior representation of a content is required for subvocal speech to be non-accidental, then prior representation of a content should be required for an

unconscious inner representing to be non-accidental. And the same reasoning would require a still earlier representation of a content; and so on.

The best thing to do would be to cut off this regress before it gets started. But what alternative view will allow us to do this? The suggestion that seems most fruitful to me is to begin by admitting our ignorance. We do not have a good theory of cognition.< 6 > We do not know how our brains make us intelligent; we do not know how our brains enable us to bring relevant knowledge to bear on situations that are new to us. We do not know how our brains manage to connect sensory materials that have arrived at widely different times with a muscular output that makes sense in the light of those earlier sensory experiences.

We do know *that* our brains do these things. And we have just seen why supposing that speech can be intelligent only if its content has already been represented in a (conscious or unconscious) inner 'speech act' is a bad start on a cognitive theory. So, let us resolutely deny such a bad start.< 7 > If we do that, we will regard speech, whether subvocal or overt, as a product of brain processes, none of which need have a content that matches that of our speech. They produce that content, but they do not have it. That kind of claim must be true at some level, and there is no reason why that level should not be the first.

Let us call this the "New Product" (NP) view. (The idea behind the name is that the content of an utterance can be a new product, i.e., one that does not need to be copied from the content of a prior event.)< 8 > NP is plainly incompatible with MT in both its forms: there is nothing in NP that has a content that an utterance could be matching. There are no "thoughts" left, if these are conceived of as they are in MT. But now, one may worry about what happens to (WT). That is, what sense are we able to make of "Our words express our thoughts" if there aren't any thoughts for our words to express?

The reading we must give to this claim if we adopt NP is this. Our words are produced by a cognitively organized brain. They express our thoughts if and only if their content fits coherently with other productions of the same brain. Malfunctions are possible, and so it is possible that we produce an utterance that does not express our thought. Slips of the tongue would be trivial examples. Confabulations are also examples, and far more interesting.< 9 > In general, however, the brain that produces an utterance of "P" produces other utterances, and is prepared to produce still more. It produces these under task demands provided by the environment, where the environment includes both surrounding objects and other speakers. In general, one's utterances fit in with the environment, with one's other utterances, and with one's actions in a coherent way. They are non-misleading reflections of the condition of one's cognitive organization. *That* is what it is for an utterance to express our thoughts.< 10 >

This account satisfies the constraint on "express", according to which it should come out true that words are arbitrary markers that would not have any meaning unless they were animated by thoughts, that is, unless they expressed what we think. And it satisfies the further constraint that it should seem *obvious* to us that our words express our thoughts.

If we adopt NP, are we thereby epiphenomenalists? No and Yes. NP is unlike traditional epiphenomenalism when it denies that there are thoughts, as familiarly conceived. For traditional epiphenomenalism does not deny that qualitative events exist. It says they exist, but are inefficacious. But we may surely say that NP is like epiphenomenalism in holding that a certain relation that is often conceived in causal terms does not have the causal structure it is often assumed to have. NP is like epiphenomenalism in denying that our (linguistic or nonlinguistic) behavior is to be conceived as the effect of a series of occurrences that have contents that would be matched by the contents of reports of what we thought.

- I V -

A corollary of NP is that knowing what we think is not, in general, getting into a state that is caused by the thing known, i.e., a state that is caused by a thought that is thereby known to be our thought. For, in general, if NP is right, we often say what we think, and know that that is what we are doing, when there is no "thought" event that has a content that "matches" the content of our (subvocal or overt) speech.

There is a reason why this point is hard to see. Namely, we *sometimes* do know what we think because we remember something we previously said (to ourselves or to others), and are simply reporting it from memory. In such cases, indeed, we know what we think because we are reporting a prior event that had a content that matches the content of our report. But the upshot of the previous section is that we cannot take cases of this kind as a general model for the relation between our intelligible words and our cognitive processes.

This observation yields a parallel between how we stand with respect to our propositional states, and how we stand with respect to our qualitative events. Namely, in both cases we know something about ourselves, and in both cases this knowledge does not rest on causation by the thing known. Or, a little more carefully: The fact that our reports of our qualitative events state something we know does not rest on causation of those reports by the reported qualitative events, and the fact that our declarations about our propositional states state something we know does not rest on causation of those declarations by "thoughts" whose content matches that of our declarations. (As noted, there are exceptions in the latter case. When we report on something we have previously said, our previous utterances have caused a memory trace that is accessed for the later report.)

We can express both points by saying that perception is not a good model either for our knowledge of qualitative events or (in general) for our knowledge of our propositional states. Or, we may say, a good model for our knowledge of nonmental things is not a good model for our knowledge of our own mentality. Now, this point runs against the grain of many habits we have as philosophers of mind. But, on the face of it, the idea that our knowledge of ourselves should be different from our knowledge of nonmental things seems to be an eminently intuitive stance. It is essentially the same intuition that is encapsulated in the phrase "homunculus fallacy" –

namely, the idea that one cannot explain cognitive abilities if cognitive abilities (e.g., perception) are already present in the alleged *explanans*.

- V -

I shall close by briefly considering a point raised by Alec Hyslop (1998). Hyslop makes many good points in support of epiphenomenalism, but holds back from acceptance because "we lose ourselves when consciousness ceases to be effective in what we choose." Hyslop is quite poignant about this loss, which I will try to express in my own words this way: If you look at yourself from an epiphenomenalist point of view, you can get into a frame of mind in which you feel you are at the mercy of your brain. It becomes possible to say to yourself, without a palpable sense of incoherence, that "My actions are all my brain's doing, *I* have nothing to do with it".

But it is not hard to see why we should not be daunted by this image of domination by a brain that strikes us as having an alien aspect, even though we know it is our own. We should not be daunted because the feeling of loss arises only from a misapplication of a model of control. To see that this is so, consider the example of driving to a grocery store. When we do this, we know in advance where we are going, and we make the car go where we want it to go. We *control* the car. If we apply this model to control of our own thinking, we will have to have a picture in which we know in advance where our thoughts should go, and make them go there. But the moment we make this feature of the "control" model explicit, it is *evident* that we have a misleading picture. We cannot *explain* how thought works by supposing that we *already* know where our thoughts should go. The whole point of a cognitive science is to explain how our thoughts arise, and it is evident that you cannot do that by supposing that they are already there, waiting to be "read off" so that we can use them to "guide" our (further) thoughts.

No doubt there is a certain anxiety that attends the thought that there is no guarantee that our brains will produce a constellation of outputs that will constitute useful actions (including, of course, intelligent speech that is relevant to ongoing projects). But the slightest reflection shows us that we most certainly lack guarantees; and a modest acquaintance with current science shows us that our actions are produced by brain processes. Recognition of these truths does not rest on accepting epiphenomenalism. Instead, they are intuitively obvious from reflection on science, and reflection on what cognitive science aims to do – namely to understand ourselves as knowing subjects. It is not natural to expect that models of ourselves in relation to other things should work as models of knowledge of ourselves as knowers. On the contrary, it is natural to expect that knowledge of our qualitative events and propositional states should have models different from that of perception, just as epiphenomenalism holds.

Notes

< 1 > In the argument of this paper, the causation of our actions by our neural apparatus replaces the more usual reference to the causal closure of the physical. The reason for this

replacement is that I believe the *intuitive* appeal of epiphenomenalism grows proportionally to detailed appreciation of neurophysiological detail. The generalization about causal closure of the physical is one that I fully accept; but its generality detracts from intuitive force.

< 2 > For elaboration and references, see discussion of Davidson's view and reactions to it in Robinson (1999).

< 3 > The warrants for this belief can be found in Robinson (1990) and Robinson (2019).

< 4 > Lormand, 1996, however, has held that we may always have subvocal speech, and do not notice it when we are speaking because it is drowned out by our audible speech. This view, however, does not follow from the premise he offers for it, and which I endorse – namely, that subvocal speech is produced by the same kinds of command events that, in other cases, lead to our audible speech. (For, it could well be that whatever opens the channel to overt speech also suppresses the mechanisms of subvocal speech.) Moreover, even if we did always have subvocal speech accompanying our overt speech, it would not follow that it causally contributes to our overt speech. (For, both might be effects of the same command events. Compare efferent copies of our head-moving commands, which prevent us from perceiving motion in the world when we turn our heads. These accompany our intentional head movements, but do not causally contribute to them.)

< 5 > One might think here of the distinctive, non-iconic phenomenology of (particular) thoughts alleged by Strawson (1994), Siewert (1998) and others. See Robinson (2005) for doubts about such views. Regarding the present paper, the same reasons I give against MT can be easily adapted to undercut the utility of this alleged phenomenology for rescuing MT.

< 6 > I do not accept all of Fodor's (2000) arguments, but I believe this work contains a sound line of thought that supports the claim that we do not have a good theory of cognition.

< 7 > In Robinson (2019) I give an entirely different argument against the view that our actions are caused by our beliefs and desires.

< 8 > I am not denying that we sometimes do first sort something out in subvocal ruminations, and then later overtly utter a conclusion that we have previously formulated. I am, however, denying that this scenario is a good model for a theory of thinking or overt speaking in general.

< 9 > See Nisbett & Wilson (1977a, 1977b).

< 10 > These ideas receive fuller expression in Robinson (1995; 2019).

References

Fodor, J. (2000) *The Mind Doesn't Work That Way*, (Cambridge, MA: MIT Press).

Hyslop, A. (1998) “Methodological Epiphenomenalism”, *Australasian Journal of Philosophy*, 76:61-70.

Lormand, E. (1996) “Nonphenomenal Consciousness”, *Noûs*, 30:242-261.

Nisbett. R. E. & Wilson, T. D. (1977a) “Telling more than we can know: verbal reports on mental processes”, *Psychological Review*, 84:231-259.

Nisbett, R. E. & Wilson, T. D. (1977b) “The halo effect: evidence for unconscious alteration of judgments”, *Journal of Personality and Social Psychology*, 35:250-256.

Papineau, D. (2002) *Thinking About Consciousness*, (Oxford: Oxford University Press).

Robinson, W. S. (1995) “Mild Realism, Causation, and Folk Psychology”, *Philosophical Psychology*, 8:167-187.

Robinson, W. S. (1999) “Epiphenomenalism”. Article in E. Zalta, ed., *Stanford Encyclopedia of Philosophy*, archive edition of March 31, 1999. I have updated this article several times, most recently in 2019. Available at <http://plato.stanford.edu/entries/epiphenomenalism>.

Robinson, W. S. (2004) *Understanding Phenomenal Consciousness*, (Cambridge: Cambridge University Press).

Robinson, W. S. (2005) “Thoughts Without Distinctive, Non-Imagistic Phenomenology”, *Philosophy and Phenomenological Research*, 70:534-561.

Robinson, W. S. (2019) *Epiphenomenal Mind: An Integrated Outlook on Sensations, Beliefs and Pleasure*, (New York and London: Routledge).

Rudd, A. (2000) “Phenomenal Judgment and Mental Causation”, *Journal of Consciousness Studies*, 7:53-66.

Siewert, C. (1998) *The Significance of Consciousness*, (Princeton: Princeton University Press).

Strawson, G. (1994) *Mental Reality*, (Cambridge, MA: MIT Press/Bradford).

Wundt, W. (1912) *An Introduction to Psychology*, translated from the second German edition by R. Pintner. (London: George Allen).