NOÛS 41:2 (2007) 318–334

Why Nothing Mental is Just in the Head

JUSTIN C. FISHER
University of British Columbia

Mental internalists hold that an individual's mental features at a given time supervene upon what is in that individual's head at that time. While many people reject mental internalism about content and justification, mental internalism is commonly accepted regarding such other mental features as rationality, emotion-types, propositional-attitude-types, moral character, and phenomenology. I construct a counter-example to mental internalism regarding all these features. My counter-example involves two creatures: a human and an alien from 'Pulse World'. These creatures' environments, behavioral dispositions and histories are such that it is intuitively clear that they are mentally quite different, even while they are, for a moment, exactly alike with respect to what's in their heads. I offer positive reasons for thinking that the case I describe is indeed possible. I then consider ways in which mental internalists might attempt to account for this case, but conclude that the only plausible option is to reject mental internalism and to adopt a particular externalist alternative—a history-oriented version of teleo-functionalism.

1. Introduction

This paper takes issue with the cluster of views that I call *mental internalism*. A mental internalist (about mental features of type T) holds that, at any given time, an individual's mental features (of type T) supervene upon what is in that individual's head at that time—i.e., whenever two individuals are indistinguishable with respect to what's in their heads, they must also be indistinguishable with respect to their mental features (of type T). Suppose there is a vat somewhere that contains a brain that, molecule for molecule, is just like your own.[1] The mental internalist (about features of type T) holds that, since the brain-in-a-vat is a duplicate of what's in your head, it must

share your mental features (of type T) as well. If you are dreamily pondering the relative merits of hot fudge sundaes and barefoot walks on the beach, then, according to the thorough-going mental internalist, your vat-bound duplicate must be doing this as well.[2]

There are various brands of mental internalism corresponding to various types of mental features that one might think 'in the head' duplicates must share. Some brands of mental internalism are quite controversial, while others are broadly accepted.

Mental internalism about *content* is especially controversial. Classic externalist arguments[3] have convinced many people that at least some aspect of mental content depends in part upon external factors like the causal chains leading to our thoughts from the things that our thoughts are about. However, a number of recent proposals[4] hold that there still is a very important notion of *narrow content* about which content internalism is true. In Putnam's (1973) familiar thought experiment, Twin Earthlings lead lives exactly parallel to ours, except that wherever our world contains H_2O, their world contains the superficially similar substance XYZ. Even though our water-thoughts pick out a different chemical substance from Twin Earthlings', narrow content theorists stress that there is also a sense in which their thoughts must have very much the same content as ours—all these thoughts mean something like *that local sort of clear drinkable fluid*, and they all bear the same sorts of rational inferential connections to other thoughts and plans of action. Classic externalist arguments like Putnam's don't establish that *all* forms of content are wide, nor do they establish that *other* mental features must be wide.

Mental internalism about *epistemic justification* is also controversial, due to its close relation to the highly controversial internalist position in contemporary epistemology.[5] Many people accept that when a responsibly formed true belief (e.g., that that thing is a barn) is brought about by a causal process that is quite unreliable (e.g., because there are many barn-facades nearby) then there is a sense in which that belief fails to be fully justified. This sense of justification must be an externalist one. However, many people also accept another sense of justification—we might call it 'narrow justification'—which requires only that an agent keep her mental affairs in appropriate order, and it is initially plausible that this notion might be satisfactorily spelled out in a fully internalist way.

One might reasonably suspect that the exceptions to mental internalism will remain limited to cases of content and justification. Classic externalist arguments have drawn upon intuitive links between content or justification and causal chains leading to internal mental states from external states of affairs. There is no obvious way in which parallel arguments might gain a foothold regarding other types of mental features like *phenomenal experiences*, *rationality*, *moral character*, *emotion-types*, or *propositional-attitude-types*, for there isn't any obvious intuitive link between these other mental features and an individual's causal relations to the surrounding world.[6] Perhaps because of

this intuitive independence, it is commonly presumed that mental internalism is true regarding these other mental features.[7] Classic externalist arguments do nothing to challenge this presumption (e.g., Putnam happily supposes that a Twin-Earthling's propositional-attitude-*types* will match those of its Earthling duplicate), and it seems unlikely that these classic externalist arguments even *could* be adapted to challenge it.

My goal is to show that this internalist presumption is deeply wrong, and that we should instead adopt some form of thorough-going mental externalism. For any of the mental features mentioned above, whether or not an agent has those features may depend, in part, upon factors not currently in her head. This entails a decisive victory for externalists in the perennial debates about content and justification, and it also firmly establishes that mental externalism is required *across the board*, a conclusion which has deep implications regarding how we should think about minds and mental features.

I begin (in section 2) by constructing a clear counter-example to all brands of mental internalism. I argue (in section 3) that this counter-example is indeed possible, and (in section 4) that it is indeed devastating for the mental internalist. I then propose (in section 5) two plausible externalist alternatives to mental internalism, and use a related case to argue (in section 6) that we should favor one of those alternatives: a history-oriented version of teleofunctionalism.

2. The Pulse World Counter-Example

My counter-example involves a creature from a planet that I call Pulse World. Pulse World orbits a pulsar, a star that spins like a super-fast lighthouse, emitting intense radiation in different directions. As a consequence, Pulse World is hit by a brief pulse of radiation 100 times each second. If any Earthling (human) were to visit Pulse World, the pulses would wreak havoc upon her neural functioning. Her brain would move from state to state in a manner that any Earthling brain scientist would characterize as an extremely irrational chain of thought. The hapless Earthling would be a raving lunatic.

Pulselings are creatures that evolved on Pulse World, and are well adapted to the regular pulses. Pulselings have brains much like ours, but their bodies are quite different. Their brains are hooked up to their bodies in such a way that, so long as they are hit by regular pulses, the resulting shifting patterns of neural activation do a wonderful job of processing neural inputs from Pulseling-sense-organs and generating the sorts of neural outputs that will prompt Pulseling-muscles to yield intelligent and well-adapted behavior. Pulselings build cars and interstate highways, and they study philosophy in Pulseling Universities. Without the regular pulses, a Pulseling's functional capacities would suffer greatly; her brain would move from neural state to neural state in a manner that is very abnormal for Pulselings, and not at all appropriate for the control of Pulseling-bodies. An unfortunate Pulseling

on Earth would be just as much a raving lunatic as would an unfortunate Earthling on Pulse World.

The first protagonist in my counter-example is Paula, a typical Pulseling, who at this moment is engaged in a perfectly normal Pulseling activity: driving her car along a Pulseling highway. (Of course, her success at this activity depends upon the presence of the regular pulses.) Let us imagine Paula in such a way that there is a great deal of behavioral, testimonial and historical evidence making it very natural to attribute the following mental features to her:

> Paula occurrently believes that she is driving her car and that the road sign she's looking at is the one announcing her exit. She's surprised her exit has come so soon. She non-occurrently believes many more things—things learned in childhood or in studying advanced mathematics at a Pulseling university. Many of Paula's beliefs are justified, and many count as knowledge. Paula's mood is one of giddy excitement. She enjoys the feel of wind through her tentacles, and the infra-red glow of Pulse World vegetation.

If asked, Paula would attribute (in Pulseling language) all these mental features to herself, as would anyone else familiar with Paula and Pulseling behavior.

Our second protagonist is Edna, an everyday Earthling, who, at this moment is engaged in a perfectly normal Earthling activity: playing saxophone. Let's imagine Edna in such a way that behavioral, testimonial and historical evidence make it very natural to attribute the following mental features to her:

> Edna occurrently believes that she is playing saxophone before an audience. She has many non-occurrent beliefs about her childhood and the things that she learned in her Earthling schools. Many of these beliefs are justified. Many count as knowledge. Edna is quite nervous. She fears she will make a mistake. Her phenomenal awareness is dominated by the sound of her saxophone. It sounds to her as though an embarrassing squeak might emerge from it at any moment.

If asked, Edna would attribute all these mental features to herself, as would anyone else familiar with Edna and Earthling behavior.

These descriptions of Edna's and Paula's respective behaviors, histories and situations make it clear that their mental lives must be radically different. Indeed, Edna and Paula differ with respect to each sort of mental feature mentioned above (content, justification, phenomenal experiences, rationality, moral character, emotion-types, and propositional-attitude-types). Any plausible theory of mentality must acknowledge these differences, or else give an extremely convincing reason to deny them.

Now, let me add one final very crucial stipulation to this case: *at this moment, Edna and Paula happen to be completely indistinguishable with respect*

to what's in their heads. (We may imagine that 'this moment' is a moment between pulses on Pulse World.) I maintain that nothing I've said above rules out this possibility. It is perfectly consistent to say that Edna's brain would do well to guide her Earthling body's saxophone playing *absent any pulses*, while Paula's momentarily indistinguishable brain would do well to guide her Pulseling body's drive down a highway *given the pulses*.

This case is a counter-example to all the brands of mental internalism mentioned above. The various brands of mental internalism entail that, since Edna and Paula are duplicates with respect to what's 'in their heads', they must also be alike in various mental respects. But, Edna and Paula clearly differ greatly in each of these mental respects. Hence, all these brands of mental internalism must be false.

3. Is This Case Really Possible?

The preceding argument rests upon two claims: (1) that it is possible that there be a case with the mechanics I described, and (2) that, given the mechanics of this case, Paula's mental features would be drastically different from Edna's. In this section, I offer positive reasons for thinking that this case is indeed possible.

I begin by discussing a related case that is itself problematic for many mental internalists. Mental internalism stakes a general claim about *all* possible agents. Many mental internalists admit the possibility of silicon-headed mental agents—e.g., future humans living in an era where it is normal to replace aging neurons with silicon nano-circuitry that (in normal human circumstances) is close enough to functionally equivalent to the original neurons. Now, silicon circuitry clearly may be such that it contains all the pathways for two very different flow-charts—one that describes an effective control system for an Earthling body, another that describes an effective control system for a Pulseling body—and such that, at numerous critical junctures, there are radio receivers that will direct current down one or the other sort of pathway, depending upon the presence or absence of a regular pulse. If Edna's and Paula's heads are both filled with silicon circuitry like this, then Edna's circuitry may do well to guide her body absent the pulses, while Paula's momentarily indistinguishable circuitry may do very well to guide her very different body given the presence of pulses. Insofar as it is clear that these two agents have very different mental features (an issue taken up in the next section), this case itself is a counter-example to mental internalism.

Still, it may be debatable whether it is even possible for silicon-headed agents to produce intelligent behavior, or whether this possibility has much bearing upon the case of actual humans living today. I will now offer reasons for thinking that, *whatever* stuff is in our heads producing our behavior—be it neurons, micro-tubules, ectoplasm, or whatever—this stuff will, in principle, be subject to the same sorts of considerations as those that apply to silicon

circuitry. Even if Edna is an ordinary present-day human, there are possible pulses[8] whose presence would make the stuff in her head—*whatever* it is—be well suited to the task of controlling some completely alien body in alien circumstances.

I should stress that I don't mean to claim that any *nomologically possible* pulse could affect human heads in this way. For all I know, it may turn out that any nomologically possible pulse powerful enough to cause the requisite changes would be too powerful for fragile human neurons to withstand. But that is beside the point. What is relevant is the *logical* or *conceptual* possibility of these pulses. With a nod to Putnam (1973), I might call them "XYZ-pulses." Just as the nomological impossibility of Putnam's XYZ was irrelevant to his argument, the (alleged) nomological impossibility of my XYZ-pulses is irrelevant to the conceptual point at issue here.

My argument for the possibility of such pulses requires three plausible presumptions. First, I presume that our behavior is produced as the consequence of a complex set of relatively simple interactions between many elements in our heads. This presumption will seem plausible to many physicalists, who might, for example, take the interacting elements to be neurons. This presumption should also be acceptable to many non-physicalists who think our behavior is produced through complex interactions involving non-physical substances and/or non-physical properties.[9]

Second, I presume that various pulses may coax the elements in our heads effectively to implement various simple interactive relations: for any given set of these elements and any given simple interactive relation that such elements might exhibit, there is a possible pulse whose effect it would be to make those elements effectively exhibit that relation. E.g., if neuron N functions as an AND-gate (by becoming highly active just in case it receives high stimulation in both its dendrites) in Edna's Earthly environment, then there is a possible pulse whose regular presence would make N effectively function as an OR-gate instead. One simple way to achieve this would be for the pulse to add effective stimulation to N, so that stimulation in just one dendrite would suffice to make N highly active. This second presumption should be broadly acceptable—for *any* quite simple effect, there is a logically possible thing that would cause it, and that thing might as well be a pulse.

Third, I presume that that there are logically possible pulses that effectively mix and match the (simpler) effects of other logically possible pulses: whenever there is one possible pulse that would serve to make the set S_1 of elements exhibit the interactive relation R_1, and another possible pulse that would serve to make some distinct set S_2 of elements exhibit interactive relation R_2 (where S_2's exhibiting R_2 is logically compatible with S_1's exhibiting R_1), there is some possible pulse that would produce both these effects. E.g., if there is a possible pulse that would make neuron N_1 effectively function as an AND-gate, and another possible pulse that would make neuron N_2 effectively function as an OR-gate, then there is a possible pulse which

would *both* make N_1 function as an AND-gate *and* make N_2 function as an OR-gate.

In the case where N_1 and N_2 are intrinsically different, this third presumption is immediately plausible—there is a logically possible pulse that would both (1) resonate with the distinguishing features of N_1 yielding the requisite effects in it, and (2) resonate in a different way with the distinguishing features of N_2 yielding the (different) requisite effects in it.

Things are less immediately obvious in the case where N_1 and N_2 are intrinsically indistinguishable. How will the pulse 'know' which of these to make work as an AND-gate, and which to make work as an OR-gate? It is fair to presume that N_1 and N_2 may be distinguished by the different *relations* that they bear to other elements in the head—e.g., N_1 might be located anterior to N_2. Now, it is logically possible for a pulse to resonate briefly in each head it reaches, in such a way that how it ends up affecting particular elements in a head will depend, in part, upon those elements' relations to other elements in that head. E.g., there is a logically possible pulse P that, upon passing through a human brain, would generate a smaller pulse Q that would itself begin at the front of that brain and sweep backwards, having different effects on the neurons it passes, depending on how far back it has gotten.[10] In this way, P might bring about quite different effects in intrinsically indistinguishable neurons N_1 and N_2. (See figure 1.)

Given our first presumption, Edna has enough neurons and enough synapses connecting them to do the work of running a Pulseling body; but in an Earthly environment, the flowing patterns of activation in her brain wouldn't be right for this task. For a brain like Edna's to do this work, activation would need to be coaxed to flow through some synapses, and not to flow through others.

By the second presumption, for each relevant bit of coaxing that might be needed, there is some logically possible pulse that would provide that bit of coaxing. And, by the third presumption, there is a logically possible pulse that mixes and matches these bits of coaxing in order to deliver all the requisite coaxing in one package. If Paula is regularly exposed to pulses like these, then a brain (momentarily) just like Edna's would do a great job at the task of controlling Paula's body. Indeed, this brain's performance at this control-task can be as robustly intelligent-seeming as you like. Hence, I conclude that my Pulse World counter-example is logically possible even when we take Edna to be an ordinary human, regardless of what complex set of interacting elements we take to be doing the important work in ordinary human heads.

We may draw a general moral from this. The normal functioning of *all* cognitive systems deeply depends upon their getting appropriate support (or at least appropriate non-interference) from their surroundings. For any complex cognitive system, there are possible surroundings in which that system would effectively perform cognitive control tasks completely different from those it

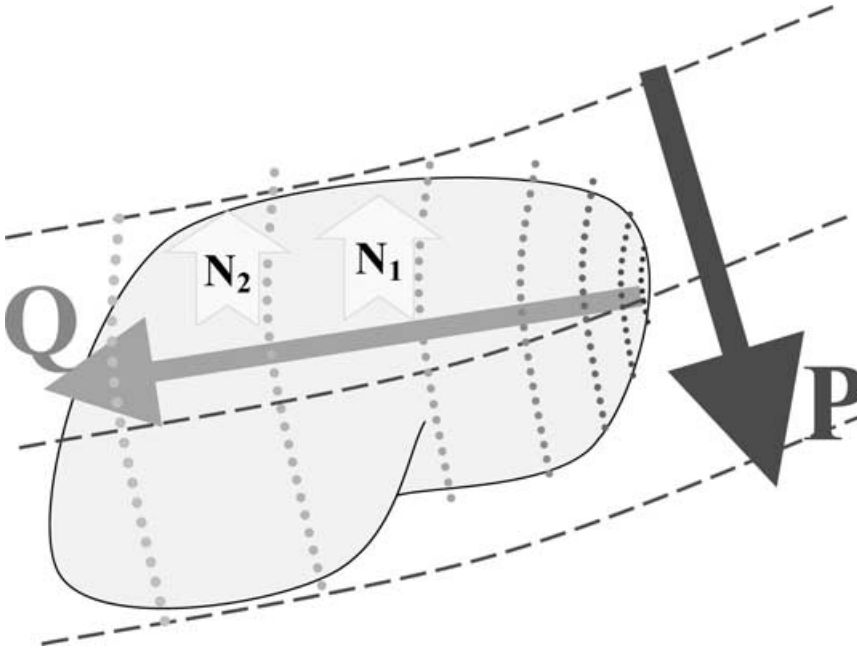

Figure 1. Various stages are depicted as Pulse P passes (downwards) through a brain and generates a smaller pulse Q at the front of the brain. Various stages are also depicted as Q sweeps backwards, having different effects on neurons N_1 and N_2 because of their different distances from the front of the brain. This illustrates how pulses like P might regularly bring about appropriately different effects even in neurons that are intrinsically indistinguishable.

normally performs. Given only the current internal features of a cognitive system, there is no way to tell which sort of control tasks (if any) this system normally performs, nor is there any way to tell which sorts of tasks (if any) it could well-perform in its current surroundings. Since the current internal features of a cognitive system can't tell us any of this, it perhaps shouldn't be surprising that (*contra* mental internalism) these internal features also can't tell us which mental features that cognitive system has.

The case of Edna and Paula provides a representative example of the general fact that the normal functioning of any cognitive system depends upon its normal surroundings. Let us now use this case to drive home the problems such dependence poses for mental internalism.

4. From Mechanics to Mentality

My argument depends not just upon the possibility of a case with the mechanics I describe, but also upon the inference from such mechanics to attributions

of quite different mental features to Edna and Paula. I will now consider potential ways of resisting this inference.

There is a classical line of thought which holds that my project is doomed from the start.[11] According to this line of thought, you have a special sort of first-person access to your own mental features, and you can never be in a position to attribute such features to any other thing, no matter how much you know about its constitution, behavioral dispositions, or history. So, for example, you can't really know that *I* have mental features, no matter how much you learn about how my brain produces my behavior; nor can you know that your pencil lacks mental features, even if you learn all about its mechanical constitution. For someone gripped by this "problem of other minds," the mechanical structure of my counter-example will seem a very poor basis for concluding that Paula the Pulseling has mental features. It is hard for me to offer arguments to this radical skeptic about other minds, just as it is hard to offer arguments to other radical skeptics. But three things are worth noting.

First, the radical skeptic about other minds has no reason to favor *mental internalism* over various sorts of mental externalism. For, she is skeptical of *any* proposed connection between mechanical features and mental features. She should be no more ready to attribute mental features like her own to a creature containing something just like her brain than she should be to attribute such features to a creature holding something just like her pencil.

Second, it is clear that the skeptic's standards are *not* the ones we normally employ in attributing mental features to other people, to animals, or to the characters we see in movies—nor are they the ones that most of us will employ if we someday make first contact with space-faring aliens. Our normal talk of mental features presupposes that robust behavioral evidence is good evidence for the presence of mental features—especially when this evidence involves (seemingly) intelligent behavior and (apparent) introspective reports of the mental features in question, and especially when it is clear that this behavior is produced by internal systems capable of responding appropriately to a wide variety of stimuli.

And third, such attributions of mental features are *explanatorily useful*. By attributing beliefs, desires, emotions, phenomenal experiences, and other mental features to various creatures we (often enough) capture real patterns in what these creatures are disposed to do. These attributions help us to understand why creatures have behaved as they have, and they help us to predict how creatures will behave in the future. This explanatory usefulness places ordinary attributions of mental states on an epistemic footing broadly comparable to that of posits made in many other scientific fields. Such a footing may not be enough to convince the radical skeptic, but it is enough for science.

So, radical skepticism about other minds flouts common usage and rejects a very useful explanatory apparatus, and it simultaneously undercuts any

reasons we might have had to accept mental internalism. In light of these drawbacks, the defender of mental internalism will probably do better to grant that various sorts of mechanical evidence—evidence about behavior, inner constitution, and history—*can* provide evidence of mentality, but to insist that the particular mechanical evidence I describe is not sufficient to sustain the conclusion I draw.

But this too is a losing battle for the mental internalist. For the mechanics of my counter-example are such that I can build in *as much mechanical evidence* as one could possibly want. Edna can be a perfectly ordinary Earthling, and hence can deserve mental state attributions as much as any of your neighbors. Pulseling behavior can be just as intelligent, just as adroit, just as eloquent, and just as well-explainable in mental terms as human behavior is—or much more so, if that would help. This means (barring radical skepticism) that we can have very strong reason to attribute mental features to Pulselings. Since Paula can be a perfectly normal Pulseling engaged in a normal activity, we can therefore have very strong reason to attribute to her mental features befitting a Pulseling in such circumstances. And I can make Edna's and Paula's situations be so different as to make the mental features we attribute to them differ in all the respects mentioned above. Hence, so long as you admit that rich mechanical evidence can constitute a good basis for attributing mental features, I can deliver however much mechanical evidence is needed to seal the counter-example.

At this point, the most promising line of defense for the mental internalist is probably to draw attention to the regular pulses that hit Paula, and to hold that the influence of these pulses somehow disqualifies Paula from receiving normal attributions of mental features on the basis of evidence about her history, constitution, and behavior. This defense seems to respect both our intuitive confidence in our attributions of mental features to Earthlings like Edna and our natural intuition that the dramatic impact of regular pulses on brains like ours *must be* highly disruptive to cognition. However, this defense comes with several prohibitively high costs.

First, this defense holds that all the Pulselings' accomplishments—their cars, highways, and philosophical writings—are products of a fundamentally non-rational process. This overlooks a level of explanation (in folk mental terms) highly relevant to understanding how these accomplishments came about. The defender is unable to make many useful and intuitively well-grounded distinctions: between the Pulseling who is content and the Pulseling who is in agony, between the Pulseling who has true beliefs and the Pulseling who believes falsehoods, between the Pulseling who makes rational choices and the Pulseling who chooses irrationally. This spectacular explanatory failure across an entire species weighs heavily against the defender's position.

Second, the defender risks saying that we humans should also be disqualified from receiving normal mental-state attributions. Many environmental

factors impinge upon and sustain our brains. It is normal for a human brain to be bathed in blood that is oxygen-rich and hallucinogen free; to be exposed to not too many G's of gravity and not too much radiation; to receive sensory stimuli that are structured in some particular ways and not in others (e.g., not in ways that cause epileptic seizures). The defender must hold that pulses are disruptive to rational cognition while a continued supply of oxygen-rich blood is not, even though both are external factors that cause neurons to behave much differently than they otherwise would have. It's hard to see how this distinction could be principled.

This problem worsens when we consider what might *for all we know* be true of our world. For all we know, our own sun emits pulses of some (heretofore undiscovered) causally active sort, and normal human cognitive processes are deeply dependent on the presence of these pulses. If it turns out that our normal behavior requires such pulses, we would rightly conclude that this just means that normal human cognition is sustained by one more external factor than we originally thought. However, our internalist defender seems committed to saying that, if it turns out that we are regularly hit by pulses like these, then we are no more deserving of mental state attributions than the Pulselings. Our own cars and highways and philosophical writings are not a testament to our rationality—instead they are just a fortuitous happenstance brought about by the regular pulses that have been jerking our poor brains around throughout human history. But clearly that would *not* be the right conclusion to draw from such evidence.

Just as we must accept that regular inputs like oxygen-rich blood or regular pulses from our sun might play an essential role in our normal cognition, we must accept that the pulsar's regular pulses play an essential role in normal Pulseling cognition. Pulselings have mental lives that are quite stable, quite rational, and quite well-suited to their ways of getting around in their environments. Despite the many features that Pulseling-brains and Earthling-brains have in common, we must understand these brains as acting as very different intelligent control systems for very different bodies in very different environments. What is normal cognitive functioning for a Pulseling's brain is quite different from what is normal cognitive functioning for an Earthling's brain. As a consequence, a Pulseling's mental features may be quite different from an Earthling's, even while their brains happen briefly to be exactly alike.

5. Two Externalist Alternatives

The Pulse World counter-example gives us compelling reason to reject all brands of mental internalism, and to accept that an individual's current mental features must depend upon *something more* than just what is currently in her head.[12] But, *how much more* must be included in the supervenience base of the mental?

I can see two plausible options.[13] First, we might take into account enough of the *current surroundings* of the individual's head to determine what sorts of ensuing processes will be likely to occur within that head. This might lead one to adopt something like existing *wide functionalist* positions.[14] Alternatively, we might take into account enough of the individual's *history* to determine what sorts of in-the-head processes are 'normal' or 'appropriate' cognitive processes for that individual. This might lead one to adopt something like existing *teleo-functionalist* positions.[15]

As far as the *above* case is concerned, we may find a difference between Edna and Paula by looking a short distance outside their heads, or a fractional second into their histories. However, other Pulse-World-style cases may be constructed to force us to acknowledge that the mental supervenience base extends further out and/or further back. There are many interesting questions here. *How far* might such cases force us to acknowledge that this supervenience base extends? What principled reasons might justify our positing particular limits on its extent?

The wide functionalist probably must allow that an individual's current mental features depend, in part, upon states of affairs *some distance away* from the intuitive boundaries of the individual's head—hence including things (like approaching pulses) which have not yet had any causal impact on what is going on within those intuitive boundaries. The teleo-functionalist will probably have to allow that an individual's current mental features might depend, in part, upon what had happened to that individual *some time ago*, or even upon what had happened to his or her evolutionary ancestors. Each of these conclusions may seem somewhat counter-intuitive, but Pulse-World-style cases show us that we must accept some conclusion like this.

6. In Favor of Teleo-Functionalism

In this final section, I argue that a variant of my Pulse World case gives us reason to prefer teleo-functionalism over wide functionalism.

This argument requires a premise which I call *the principle of mental inertia*. This principle says that altering things outside a creature's head won't significantly[16] change the progression of mental states that that creature will undergo, unless those external alterations also bring about changes within the creature's head. As an extreme example, imagine a creature whose brain is quickly extracted from its body and placed in a nutrient-filled vat. There is a strong inclination to say that this surgical procedure, *by itself*, does not significantly change the creature's mental features from what they otherwise would have been. If the envatted brain receives different inputs than it would have received had it continued to be embodied, these differences *likely will* bring about mental differences. But until these different inputs have their internal effects, the brain will undergo roughly the same progression of mental states it would have undergone had it remained embodied. Intuitively, normal

brains have a sort of *mental inertia* which allows them to retain their normal progression of mental states, even while their surroundings change drastically.

The principle of mental inertia states a constraint on how creatures' mental features might change with the passage of time. This distinguishes it from the synchronic claim made by the mental internalist. Still, it is dangerously easy to suppose that these are more closely related than they are, as in the following intuitive argument:

> Surely you can't change my mental states *just* by changing things outside my head. So, my mental states must depend only upon what's in my head.

This intuitive argument is fallacious because its premise doesn't rule out the possibility that one's mental states also depend in part upon one's history. The internalist conclusion of this intuitive argument is false, as my Pulse World case shows. But its premise, the principle of mental inertia, is quite plausibly true. The internalist who was lured in by this fallacious argument might embrace the principle of mental inertia as a nugget of truth underlying her intuitive attraction to mental internalism.

It is outside the scope of this paper to argue that we *have to* accept the principle of mental inertia. For present purposes, it is enough to have shown that we have a strong intuitive commitment to this principle—a commitment that is arguably stronger and more fundamental than our intuitive commitment to mental internalism. This is enough to place an argumentative burden on the wide functionalist who must reject this principle.

To see how the principle of mental inertia causes problems for the wide functionalist, consider the case in which Edna's and Paula's momentarily-indistinguishable brains are both very quickly (within the period of time between successive pulses) extracted from their respective bodies and emplaced in indistinguishable vats in indistinguishable surroundings. The principle of mental inertia entails that this procedure, by itself, does not significantly change their mental features from what they otherwise would have been: Edna still has Earthling-playing-a-saxophone mental features, and Paula still has Pulseling-driving-a-pulsemobile mental features. (Of course, since their new surroundings are exactly alike, at least one of them is about to receive what are for her extremely abnormal inputs—and hence is about to undergo abnormal mental changes—but that hasn't happened yet.) Edna's and Paula's envatted brains are indistinguishable and they have indistinguishable surroundings; yet they have very different mental features. Hence, this case is a counter-example to wide functionalism.

What's more, it seems clear that these brains have different mental features *because* one is an Earthling brain plucked from an Earthling body, while the other is a Pulseling brain plucked from a Pulseling body. This suggests that the history-oriented teleo-functional approach is on the right track, while wide functionalism was not.

In brief conclusion, the Pulse World counter-example shows that we must reject mental internalism and allow that an agent's mental features depend upon something more than just what is currently in her head. Considerations involving mental inertia suggest that this 'something more' must lie in an agent's history, rather than in the current surroundings of her head. This gives us reason to favor history-oriented teleo-functionalism over both mental internalism and wide functionalism. Of course, there are many questions regarding *which* historical factors are the ones that matter. As we seek answers to these questions, Pulse-World-style cases will help us to dispel widely-held but mistaken internalist presumptions, and to make clear important issues that we must consider as we develop better theories of mentality.[17]

Notes

[1] If you think there is more 'in your head' than just your brain—e.g., perhaps you believe in immaterial thinking substances—then let the vat-bound being also duplicate you in these other respects as well. And if you think that important parts of cognition are performed in your spinal cord, stomach, or heart, let us count these organs as being 'in your head', too, and suppose that your vat-bound duplicate shares them as well.

[2] Mental internalism is related to *psychological individualism*—the much-discussed view that psychological taxonomy should be done only on the basis of each individual's intrinsic features. If respectable psychologists may introduce technical non-mentalistic notions, then psychological individualism stakes a much broader claim than does mental internalism. However, insofar as our folk mental notions overlap with the notions employed in a respectable scientific psychology, exceptions to mental internalism will also be exceptions to psychological individualism. (Eliminativists may deny that there is such an overlap; others may deny that the overlap includes such notions as rationality, justification, or phenomenology.)

[3] Classic arguments for content externalism include Kripke (1972), Putnam (1973), and Burge (1979).

[4] See, e.g., Chalmers (2002), Dennett (1981), Fodor (1987), Loar (1988), Segal (2000), and White (1982).

[5] Classic arguments for epistemological externalism include Dretske (1981), Goldman (1986), and Plantinga (1993). Epistemologists have used the term 'internalism' in various ways—sometimes meaning that justification depends upon what's 'in the head', but often meaning only that justification depends upon *what's in conscious experience*. If one holds (as I think one should hold) that conscious experience doesn't supervene upon what's currently in the head, then one might conceivably retain this latter sort of 'epistemic internalism' without thereby committing oneself to what I call 'mental internalism about justification'.

[6] One semi-plausible strategy would be to sketch intuitive links between other mental features and *content*, and then to argue that the relevant sort of content depends upon external causal chains. The rationality and moral appropriateness of various 'in the head' changes plausibly depend upon the *content* of one's mental states. Similarly, it is arguable that one's phenomenal states typically have content (see, e.g., Horgan & Tienson 2002, Siewart 1998, Tye 2002, Dretske 1995, or Dennett 1991), and that the types of one's emotional states are determined at least in part by their content (for a useful survey see Prinz 2004, pp. 7–9). However, it is not at all clear that the sort of content involved in these cases should be the same as the sort of content that Putnam/Burge arguments show to be wide. And, regardless, there is no obvious way to generalize this to the case of propositional-attitude-types.

[7] There are a number of notable challenges to this presumption, including Millikan (1984), Dretske (1995), Lycan (2001) and Tye (2002). Tye writes, "Internal supervenience for the phenomenal is no more than a dogma. And sleeping dogmas should not be left undisturbed." (Tye 2002, pg. 453). I hope this paper will do some disturbing.

[8] There may be *non*-pulse-like external influences that would do this too. I will stick with pulses though, because they are relatively simple and easy to think about, and because they allow for a satisfyingly non-instantaneous period of time (between pulses) throughout which Edna and Paula remain 'in the head' duplicates.

[9] One might want to weaken this presumption to admit for quite complex or holistic patterns of interaction. I think my general conclusion still holds in such cases, but it is harder to make a straight-forward argument for this.

[10] If we could expect that the brain will always retain the same distance and orientation with respect to the source of Pulse P, and that N_2 will always be further away from this source than N_1, then we could easily do this without talking about Pulse Q. But, to allow the possibility that Pulselings might move around freely (and do somersaults) on Pulse World, it is useful to have shown how a logically possible pulse might differentially affect neurons on the basis of their respective relations to *other elements in the head*, rather than on the basis of their respective relations to any external landmarks like the pulsar. Pulse Q provides a straight-forward mechanism for doing this.

[11] This line of thought is present not only in the classic "problem of other minds," but also in contemporary discussions of the possibility of zombies and spectral inversion.

[12] There may be *independent* reasons to think that mental internalism must be rejected or revised. For one, mental internalism is formulated in terms of an individual's 'in the head' features at a given time. According to the special theory of relativity, there is no privileged notion of simultaneity, and hence no privileged answer to the question of which features are instantiated 'at the same time'. The mental internalist might circumvent this problem by relativizing mental features to particular frames of reference. However, it violates the intuitive spirit of mental internalism to suppose that an agent has numerous different mental features relative to numerous different frames of reference; and it is hard to see what principled reason we might have to privilege some particular frame of reference.

Second, many people who are attracted to mental internalism might hold that indistinguishable brains are guaranteed to be mentally indistinguishable *only when the brains are subject to the same universal natural laws*. One might revise mental internalism to ensure that universal natural laws are admitted into the supervenience base for the mental. But this revision doesn't help against my counter-example, as Edna and Paula may dwell within the same universe and be subject to the same universal natural laws.

[13] These do not exhaust the logical possibilities. For example, one might consider various hybrids of the two options I propose. Alternatively, one might follow Shagrir (2001), who in considering a somewhat related case, suggests employing an externalist theory of content to determine which contents to attribute to various 'in the head' states, and hopes that this might adequately constrain our assessment of what might count as the functioning going on in that 'head'. However, it's not clear how one might apply a theory of content independently of a theory of functioning; nor that attributions of *wide content* could adequately constrain possible attributions of propositional-attitude-types, emotion-types or phenomenology.

[14] See Harman (1988), Patricia Kitcher (1991), and Wilson (1995).

[15] See Millikan (1984), Neander (1991), Dretske (1991), Papineau (1993), and Price (2001).

[16] There may be some mental features that do change as surroundings change. E.g., some defenders of wide content will hold that the wide content of perceptual-demonstrative thoughts includes *whatever thing it is* that is located in an appropriate location relative to the perceiver—such wide content would change if surroundings change. Changes in surroundings may also alter the truth of many justified beliefs, and hence stop them from counting as *knowledge*. Although it is hard to give a principled delimitation of the set of mental features that are

'super-context-sensitive' in this way, we are intuitively committed to the claim that many mental features are not so 'super-context-sensitive', and instead have enough inertia to survive changes in surroundings (at least for a while). In the discussion that follows, I will bracket 'super-context-sensitive' mental features, and concentrate upon the more ordinary, less context-sensitive ones.

[17] This work was supported by NSF grant no. IRI-IIS-0080888. I am especially indebted to David Chalmers for many useful comments upon multiple drafts of this paper. I am also thankful for helpful comments from Stephen Biggs, Doug Campbell, Travis Fisher, Terry Horgan, Ruth Millikan, David Slutsky, Sarah Wright, two anonymous reviewers, and audiences at the University of Arizona, the Association for the Scientific Study of Consciousness, and the Society for Philosophy and Psychology.

References

Burge, Tyler. (1979) "Individualism and the Mental," in *Studies in Metaphysics*. P. French, T. Uehling, and H. Wettstein, eds. Minneapolis: University of Minnesota Press.

Chalmers, David. (2002) "The Components of Content," in *Philosophy of Mind: Classical and Contemporary Readings*. David J. Chalmers, ed. New York: Oxford University Press, pp. 608–33.

Dennett, Daniel. (1981) "Beyond Belief," in *Thought and Object*. A. Woodfield, ed. New York: Oxford University Press.

——. (1991) *Consciousness Explained*. Boston: Little, Brown.

Dretske, Fred. (1981) *Knowledge and the Flow of Information*. Cambridge, MA: Bradford Books, MIT Press.

——. (1991) *Explaining Behavior*. Cambridge, MA: Cambridge University Press.

——. (1995) *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Fodor, Jerry. (1987) *Psychosemantics*. Cambridge, MA: MIT Press.

Goldman, Alvin. (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Harman, Gilbert. (1988) "Wide Functionalism," in *Cognition and Representation*. Stephen Schiffer and Diane Steels, eds. Boulder, CO: Westview Press, pp. 1–12.

Horgan, Terry, and John Tienson. (2002) "The Intentionality of Phenomenology and the Phenomenology of Intentionality," in *Philosophy of Mind: Classical and Contemporary Readings*. David J. Chalmers, ed. New York: Oxford University Press, pp 520–33.

Kitcher, Patricia. (1991) "Narrow Psychology and Wide Functionalism," in *The Philosophy of Science*. Richard Boyd, Philip Gasker, and J.D. Trout, eds. Cambridge, MA: MIT Press, pp. 671–85.

Kripke, Saul. (1972) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Loar, Brian. (1988) "Social Content and Psychological Content," in *Contents of Thought*. R.H. Grimm and D.D. Merrill, eds. Tucson: University of Arizona Press.

Lycan, William. (2001) "The Case for Phenomenal Externalism," *Philosophical Perspectives*, *Vol. 15: Metaphysics*. Atascadero: Ridgeview Publishing, 2001.

Millikan, Ruth. (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Neander, Karen. (1991) "The Teleological Notion of Function," *Australasian Journal of Philosophy* 69, pp. 454–68.

Papineau, David. (1993) *Philosophical Naturalism*. Oxford: Basil Blackwell.

Plantinga, Alvin. (1993) *Warrant and Proper Function*. New York: Oxford University Press.

Price, Carolyn. (2001) *Functions in Mind: A Theory of Intentional Content*. New York: Oxford University Press.

Prinz, Jesse. (2004) *Gut Reactions*. New York: Oxford University Press.

Putnam, Hilary. (1973) "Meaning and Reference," in *The Philosophy of Language*. A.P. Martinich, ed. New York: Oxford University Press, 1996.

Segal, Gabriel. (2000) *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.
Shagrir, Oron. (2001) "Content, Computation and Externalism," *Mind*. Vol. 110, Iss. 438, pp. 369–400.
Siewart, Charles. (1998) *The Significance of Consciousness*. Princeton: Princeton University Press.
Tye, Michael. (2002) "Visual Qualia and Visual Content Revisited," in *Philosophy of Mind: Classical and Contemporary Readings*. David J. Chalmers, ed. New York: Oxford University Press, pp. 447–56.
White, Stephen. (1982) "Partial Character and the Language of Thought," *Pacific Philosophical Quarterly* 63:347–65.
Wilson, Robert. (1994) "Wide Computationalism," *Mind*. New Series. Vol. 103. Iss. 411, pp. 351–72.