# Some puzzles concerning relations between minds, brains, and bodies[1]

Rick Grush

Department of Philosophy 0119
University of California, San Diego
La Jolla, CA 92093

ABSTRACT: In this article I explore a number of questions that have not been adequately investigated in philosophy of mind circles: are minds located in the same place as the brains (or other computing machinery) supporting them? Must they exist at the same location as the body? Must they exist at the same time? Could a single mind be implemented in multiple brains, or multiple minds in a single brain? Under what conditions might a single mind persist despite being implemented successively in different brains? What contributions do features of the computing machinery make to these questions, compared to the contribution made by the body and embedded point of view? Some of these questions have been touched on previously, but there hasn't been any attempt at a systematic analysis of the various consequences that different approaches in the philosophy of mind have for how the spatiotemporal location, synchronic individuation and diachronic identity of minds relates to the spatiotemporal location, synchronic individuation, and diachronic identity of both the implementing computational machinery and the embodied embedded point of view. I make a first stab at such an analysis by discussing a variety of thought experiments in which such questions of location, individuation, and identity arise, and I explore how various approaches to understanding the mind – identity theoretic, functionalist, contentualist, embodied/embedded/extended, and so forth – would respond to such situations. A number of novel issues emerge, and some surprising affinities are revealed.

## 0. Introduction

One of the main issues occupying philosophers of mind is the relationship between the mind and brain. Recently, the role of the body and environment in this relationship has gathered significant interest. But the basic problems are the same: What are minds? What are their physical bases? And what is the relation between minds and their physical bases (whatever they are)? Most views fall into one of three very broad categories: the mind is *identical to* its physical basis; the mind is (functionally or computationally) *implemented in* its physical basis; or the mind is some sort of content-defined virtual or abstract entity produced or *constructed by* its physical basis. Shortly, I will characterize these three broad views in a little more detail. In what follows I will use the expressions 'mind', 'self', and 'subject' more or less synonymously, and I will use the expression 'implementation' and its cognates as general-purpose relation terms intended to cover all of the three above sorts of relations and anything of the same general sort.

A mainstay of philosophical exploration of such views is *substitution of the basis*. A thought experiment is produced in which something other than the usual basis is substituted for the usual basis, and intuitions about what would result are canvassed. These results in turn are taken to have consequences for this or that theory. Consider a few examples from this long history, all of which involve taking the normal human brain as the prototypical basis of mentality and riffing on various substitutions: Martian Brains (Lewis 1980), which are supposed to be able to implement minds, or mental states, and hence demonstrate that the type-identity theory is too strict; the Chinese Room (Searle 1980), which is meant to show that implementing the same input/output function (or even potentially the same internal functional system) as the brain is insufficient for mindedness, and hence certain kinds of artificial intelligence approaches can't be correct; the Nation of China and Blockhead (Block 1980), in which the functional structure of a human brain is duplicated by the population of China interacting in certain ways, which is supposed to pose a challenge for functionalism; the replacement (or supplementation) of certain parts of the brain with environmental tools or props (Clark and Chalmers 1998), which is intended to show that the mind can be

extended out to the environment and hence that the brain itself is not the proprietary basis of the mind; and so forth.

My goal is not to add to this list. Rather, I want to start a new list. In the following six sections I will be exploring some manipulations on the implementation bases of minds. But the manipulations won't involve *substitutions* of different bases. In all cases the manipulations will involve only biological brains, normal bodies, and normal environments – no substitutions of neurons by silicon chips, nations-of-China computing functions, or anything of the sort. That is, the manipulations will not involve any *substitutions*. Rather, the bases will be manipulated in other ways, and the investigation will hinge on what various theories would say about these cases, as well as what intuition might say about them. In Sections 1, 2, and 3, the manipulations will be spatial and temporal, and the effects on the spatial and temporal features of the mind will be explored. In Sections 4, 5, and 6, the impact of manipulating identity and individuation features of elements of the basis on the identity and individuation features of implemented minds will be explored. These manipulations raise a number of novel issues for views of mindedness in general, and they also raise novel challenges for most if not all accounts of the relation between minds and their physical bases.

Let me say a bit more about the three broad views I will use as stalking horses throughout. The first is the position that the mind just *is* the brain — henceforth *brain-lubber*. This covers a diverse collection of philosophers, from identity theorists to Searle – who argues that consciousness is a product of causal powers of the biological brain (Searle 1980) – to some property dualists, as well as elements of those positions that claim that there is something special about a naturally evolved brain. What unifies these positions is the idea that there is something *quite special* about the *matter* of the brain, maybe even the naturally evolved brain, itself. It might even be that what is special about it is that any given brain has some proprietary connection to some specific non-physical thing that enables mentality or some mental features.[2] There is no guarantee that anything other than *my* brain could implement *my* mind even if it processed information exactly as my brain does or, more generally, implemented exactly the same functional, causal, or computational organization that my brain does.

[2] It might seem odd that I include dualism in the brain lubber camp. But for the purposes of the sorts of thought experiments that will be the topic of this essay, they are in the same category. After the reader has made it through the first few sections it will be clear why this is the case.

And perhaps even a molecule-for-molecule replica, not produced by natural selection, would not have mentality. Prototypical brain-lubbers will agree that the functional features of the brain and its computational powers are important. But they insist that they are not sufficient for mindedness, the actual biological stuff is also necessary.

There is a position akin to the brain-lubber's, which has it that it is the entire living body, not just the brain, that is the seat of mentality. And perhaps all or parts of the environment as well. I'll refer to these views as *brain/body/environment-lubbers,* or BBE-lubbers for short. Like brain-lubbers, they recognize that the functions and even computations or information processing are crucial, but they generally insist that the specific material implementation is also necessary.

The second broad approach, after lubberism, is *functionalism*. This is the view that *mental states* are functional states, and by extension a *mind* is a functional organization capable of implementing a suitable complex of such states. *Functionalism*, as I am using the term, is also much wider than it might first seem. Anyone who takes it that it is the computational or information-processing capabilities that a brain has that yield mentality (apart from any special physical implementation) would count as a functionalist.

Functionalists may differ with respect to what they think it is that is functionally important for mindedness. Supposing that we understand what the functional specification sufficient for mindedness is, some functionalists will take it that the brain is the typical implementor, others may include the body, and yet others may well include the environment (e.g. extended mind proponents). The options here are parallel to the options available for the lubbers. The difference is that for the lubbers, it is the brain itself, or the brain and body themselves as biological entities that is crucial for mindedness. For the functionalists, it is a distinction of what the normal physical implementors of the relevant functions are. Some proponents of embodied cognition will claim that it is the actual biological body that is crucial; others – for example those who emphasize the dynamics of the body – may well agree that the body is important, but only because it fills certain important functional or causal/dynamical roles.

Note that the distinction between wide and narrow content (Block 1980) is not the same as the distinction between functionalists who think that the brain is the typical implementor of the relevant function and those who think that the body and environment are also implementors. Even one who thinks that the relevant functions

are all implemented in the brain might insist that external objects are part of the individuation of any contents *carried by* those states.

The final approach is neither lubberist or functionalist. I will call it *contentualism*. The name isn't a marketing homerun, but it seems to me better than the other options I could think of. The main idea is that a mind or self is a certain kind of organized system of content attitudes, typically a model, virtual entity, or narrative of some sort.[3] A familiar example would be Dennett's theory of selves, according to which a self is a virtual content-defined entity, a "center of narrative gravity". A narrative is a certain kind of set of appropriately structured contents, and so it is an exemplar of the contentualist approach (Dennett 1992). But this camp is large. Another example would be Robert Brandom's (1994) view, according to which a person is a nexus of discursive commitments and entitlements structured according to norms of deontic scorekeeping. It includes Ismael's (2007) view of selves as certain kinds of dynamical self models, Metzinger's transparent self-models (2003), and so forth. Models, narratives, nexus of discursive commitment and entitlement, and the like are all content-structures, and views that place such structures at the core of what qualifies as a mind or self will count as contentualist. Of course, just because such content structures are in a sense abstract, it does not follow that they are causally inert (see Ismael 2007, 2011).

One might think that contentualism is a special kind of functionalism. The suggestion would be that what is special about it is that the functional system has the specific job of creating and maintaining a relevant content structure. But this isn't quite right. The issue is *what is* the self/mind. A contentualist might agree that an appropriate content structure can only be produced by a certain sort of functional system. But the mind or self that results would not *be* the functional system, but rather the constructed content structure. The distinction is important. One might be able to construct a map with some sort of functional organization of blocks, for example. And in such a case, the functional system of the blocks that explains how it can implement a map are one thing, but the terrain mapped, the thing the map is about, is another. In short, the content/vehicle distinction is still important when the vehicle is a functional system.

[3] I say "content attitude" rather than "propositional attitude" simply because I don't want to get needlessly mired in debates about non-conceptual or non-propositional content.

When convenient, I will use Dennett's account as a placeholder for the contentualist category, since it is probably the most familiar. The differences, though substantial, between it and the other members of the category won't really matter until Sections 5 and 6. When I wish to emphasize that it is features of the general category that are under discussion, I will refer not to narratives but to *content-structures*. This is a generic term that I intend to cover Dennettian narratives, Brandomian nexus of discursive commitments and entitlements, Ismaelian dynamical self/world models, and the like.

There are two relevant aspects of the content-structures appealed to by contentualists. On the one hand, there are the specific *contents* involved in the structure. These can include propositions such as "I was born in Oregon", "Cats are mammals", and "That bottle up ahead is full of water". They might also include contents involved in modeling the agent and its environment, placing oneself on a cognitive map, or even representations of one's own representations. On the other hand, there is the *structure*. On Dennett's view, this would amount to the discursive framework – the narrative – in which the propositional contents are embedded; for Brandomian nexus, this amounts to inferential relations among the various discursive commitments; on a cognitive map, it would involve the represented spatial relations between represented objects. And so forth for the other contentualists.[4]

The contents involved might be individuated either narrowly or broadly. It is worth saying a bit about this distinction, since it will figure in much of what follows. My own beliefs that I would express with the word 'water' differ in a key respect from the beliefs that my molecule-for-molecule duplicate on Twin Earth (for convenience, we can call him 'Rico', though of course he would himself answer to the same phonological string as I would) would express using the same phonologically and orthographically individuated word. My belief is about $H_2O$, while Rico's is about XYZ – the stuff on Twin Earth that plays the same role as water, despite its different chemical structure.

[4] On Brandom's view, of course, (i) and (ii) won't be independent, as the individuation of contents hangs in part on the inferential structures in which those contents are embedded. Furthermore, Brandom's view is complicated by the fact that all such elements are products of attribution by members of a language community. But of course, less saliently, the same is true of Dennett's position. These latter complexities will become important only later on.

On the other hand, there is a similarity between my 'water'-expressible beliefs and those of Rico, namely, they play analogous roles in our psychology. The beliefs that Rico and I would express by saying "water quenches thirst" play analogous roles in our deliberations about addressing thirst. Narrow-content is what captures what is psychologically analogous in Rico and me. They are individuated by (short-arm) functional role, or inferential role, or something of the sort. Broad contents are those whose individuation includes reference to referents. My 'water'-sentences and Rico's 'water'-sentences express beliefs with the same narrow content, but distinct broad contents, because mine are about water, while Rico's are about XYZ. This distinction will be significant in much of what follows.

With these preliminaries in hand, we can turn to the first of the manipulations. It is familiar – it will be the only one that is familiar – and so it will be a convenient gateway for the others.

## 1. Spatial displacement of the basis: *Airhead*

In this section I will introduce the first manipulation, one first explored by Dennett (1981). I will provide more detail to the thought experiment than Dennett originally did, bring out some important features, and head off a few potential worries. I will then describe how each of the three major approaches (lubberist, functionalist, and contentualist) will respond to this situation, and along the way a number of resources will be developed that will be useful in subsequent sections.

The manipulation is the spatial displacement of the brain from the body, and the issue raised is how the location of the self would be affected by that manipulation. Dennett constructs a thought experiment in which his brain is removed from his body and put in a vat, but the neurally mediated interactions between the brain and body are maintained by tiny radio transmitters and receivers spliced onto the synapses that were separated when the brain was removed. Removal of the brain from its normal neural input and output channels would require the severing of a relatively small number of tracts, such as the spinal cord, optic nerves, and cranial nerves. Each of these tracts, of course, contains many individual neural transmission lines. The idea is that each of these tracts is severed, but all the neural signal conduction that would normally go through these tracts is kept intact by placing transmitters on the

presynaptic cells on one side, such that whenever those cells fire, the neurotransmitter released is detected, and as a result, a signal is sent to a receiver placed on the relevant dendrite(s) of the post-synaptic cell on the other side. When signaled, this receiver releases neurotransmitter.

Though the cells are no longer physically adjoining, *the causal connection is maintained*. This means, among other things, that the dynamical equations that might be used to describe the relations between the brain and body won't change. When the presynaptic cell releases neurotransmitter, the postsynaptic cell uptakes neurotransmitter, just as in the normal case. The only difference is that there is now a reliable causal intermediary. And since this is done for every signal-carrying conduit in the tract, the fact that the *tract* is no longer biologically continuous does not affect the signaling. The tracts are causally intact, and that is what matters. And by hypothesis this process happens quickly enough that normal information processing is not affected.

The thought experiment assumes (quite plausibly, in my view) that the physical proximity of neurons to the bodily elements they interact with is important only insofar as that proximity enables convenient causal exchange. The thought experiment as stated focuses on neurally mediated causal interactions, but any other causal interactions could be duplicated. If you think that body temperature is important, then we can equip the body with sensors that measure its temperature, and that signal will be sent to the vat to alter the vat's temperature accordingly, so that the brain will be in a thermal environment that is equivalent to the thermal environment it would have been in if it were in the body. If the body spends enough time in the hot desert, the person might pass out. Ditto for anything else you think is important.[5] If that causal interaction can be maintained without physical proximity – and causal interactions are the kind of things that can generally be extended beyond physical proximity – then no

[5] Cosmelli and Thompson (2010) pitch arguments against the idea of a normally functioning envatted brain, and they mention the Dennett scenario specifically. They do make a good case to the effect that the vat would have to be far more sophisticated than you might have thought. And sophisticated in the sense that it would need to replicate a lot of things that the body normally does. But their arguments against the possibility of a normally functioning evatted brain, as far as I can tell, are only a threat to the idea of an envatted brain bereft of causal interaction with the actual body. Nothing they say imperils in any way the coherence of an envatted brain in which all casual interactions between it and its actual body are maintained. They do indeed suggest that their arguments establish this result, but as far as I can tell they don't even clearly address it. The dialectic seems to establish that any such vat would have to be very sophisticated and that an envatted brain bereft of normal causal interaction with the body would not function normally. Then there is a slide from this to the conclusion that an envatted brain that maintains all casual contact with the body could function normally is not possible.

harm and no foul.[6] And of course, chemical and other influences can be replicated as well (see remarks below about alcohol).

The procedure is essentially what telephones did for verbal conversations. They added a causal intermediary between the acoustic signals leaving the speaker's mouth and those entering the listener's ear. I will call this procedure *interfacing* – the creation of interfaces (e.g. clusters of transceivers) that insert reliable but spatially flexible causal intermediaries between systems whose normal causal interaction proceeds without such intermediaries. And they do it in such a way as to have no impact on the functioning of the components. Of course telephones do have a lot of impact on the causal function. Acoustic features are compressed and changed, for instance. Telephones are very imperfect interfaces.

I want to draw attention to the impressive list of important things that are *not* imperiled or eliminated by interfacing:

1. All relevant[7] causal patterns and functional relations are maintained. Any neural setup according to which brain or body part X influences brain or body part Y in conditions Z will continue to have that causal profile after interfacing. This means, in particular, that ***everything relevant to proponents of embodied, enactive, embedded cognition is likewise preserved***.[8]

2. Material influences can be replicated as well. If the infusion of some chemical from the body to the brain, or vice versa, is important, then this can be

[6] There is no profit in squabbling over the complexity and precise details of the procedure or of the causal connections. To be sure, splicing a trillion tiny transceivers on to synapses by hand would be a chore. My guess is that in a few thousand years we'll have drinkable nanobot-infused beverages that will do this in a matter of minutes. And they'll come in strawberry, chocolate, and vanilla. And yes, perhaps it is not just the fact of two neurons synapsing that is important but also the details of the precise location of the synapses, perhaps even relative to other dendrites. And perhaps there are important causal influences other than neurotransmitter. Maybe glial cells do important work, or sub-spike potential shifts, or nitrous oxide concentrations. The key idea, though, is that *whatever* the relevant causal influences are in the case of undetached brains, the interfacing replicates them in the thought experiment. Unless there is something magical or spiritual going on, these are physical effects that can in principle be physically detected and physically reproduced. The naive belief that there are special unreplicable features of this or that physical system – including high degrees of complexity – is precisely the belief exploited by hackers and thieves when they defeat fingerprint scanners, debit card readers, mechanical locks, network passwords, and so forth.

[7] Meaning that, whatever causal interaction you think is relevant, we include that on the list of causal chains that get extended and maintained.

[8] A handful of people who've commented on drafts of this paper have been unclear on this. Of course there is a good deal of important work demonstrating that the body and brain have rich interactions, and that perhaps these interactions are more significant for brain functioning than might have been thought. This thought experiment does not deny any of this. In the situation we are considering, all of these rich causal interactions between the brain and body are maintained. The ability of the envatted brain to perform visual detection tasks might even be influenced by the body's heartbeat, for example (see Park et al. 2014).

replicated. For instance, one can put alcohol detectors in the body's bloodstream such that when they detect the presence of alcohol, a mechanism in the vat releases alcohol into the vat (in whatever way would be appropriate for replicating the normal effect).

3. The causal genealogy of afferent states and the causal legacy of efferent states is preserved. Perceptual states are still caused by the external objects interacting with the embedded body's sense organs. Proponents of causal or object-involving theories of content will have no legitimate grounds for a beef.

4. My brain is unmolested in the sense that there are no non-biological substitutions in play – no silicon neurons replacing biological neurons, or what-have-you.

5. Everything relevant to teleo/biosemantics is still in place. The envatted and interfaced brain is just as much a product of evolution as it was pre-envattment.

Given this impressive list, I will adopt two hypotheses as operating assumptions. The first I will call the *mind-preservation hypothesis (MPH)*: interfacing preserves mindedness. As far as I can tell, anything anyone has ever claimed to be important for genuine mentality is preserved through interfacing, as itemized in the list above.

There are only two things not preserved: spatial contiguity, and the *exact* casual chains, meaning the causal chains are all there, but some just have an additional link spliced in. As for spatial contiguity, I don't think I've ever heard anyone lobby that spatial contiguity (considered apart from causal and chemical interactions, of course) plays any important role.[9] As for extended causal chains being involved, again I can't recall this being an issue for anyone. Indeed, a good deal of recent empirical work on neural implants, neural plasticity, and sensory substitution (to name a few topics) point, individually and jointly, in the direction of there being a great deal of latitude in the precise details of various causal chains without threat to the associated psychological phenomena.

[9] The only possibility I can come up with would be a theory that takes quantum mechanical phenomena, such as entanglement, to be somehow involved in mentality. These are relations that aren't exactly causal and would not *necessarily* be replicable over interposed distances. Such approaches strike me as too hare-brained to merit inclusion in this discourse. For any reader who might be a proponent of such a view, don't fret over my dismissal – if you're right, eventually you'll have the last laugh.

Given everything listed above that is preserved by interfacing, it isn't clear what considerations could be mobilized to support resistance to MPH.

The second hypothesis is weaker: *the behavior-preservation hypothesis (BPH)*: interfacing preserves overt behavior. This hypothesis really hinges only on (1) above. Maybe (2) also. In order to deny BPH, one would have to deny that overt behavior, in all its physical details, is determined by physical causal structures, since all these causal structures are, *ex hypothesi*, maintained. How one might deny BPH without incorporating the paranormal is unclear to me. MPH, as I mean it, entails BPH, and so when, as the discussion proceeds, I point out that MPH is satisfied, that should be taken to imply that both are satisfied. BPH won't play a role in the dialectic until Section 6, where situations in which the possibility that BPH but not MPH holds will be raised (well, a little bit in Section 3 too). BPH without MPH is the specter of philosophical zombies, of course.

Enough by way of preliminary. Let's call the subject who has undergone this procedure – whose brain is no longer in their skull – *Airhead*. Just to put this in the first person, suppose that I am Airhead. I presume, in accordance with MPH, that from my subjective viewpoint things would be pretty much exactly how things are now. Perhaps drinking alcohol wouldn't make me drunk. The alcohol in my body's blood stream wouldn't make it to my brain, as Dennett observed, unless that connection were also replicated (as per (2) above). And I'd be more or less immune to the more serious negative effects of a rear naked choke, since the vat would be supplying my brain with oxygen and other nutritive needs independent of pressure on my carotid arteries. I would have an advantage in jiujutsu competitions – again, unless that causal chain were extended as well.

But aside from minor benefits like that, I see no reason to think that my daily life would be any different. I would be seeing through my eyes as I do now, hearing through my ears, and moving my arms and legs as I do now. It might even take some convincing to get me to believe that the procedure had actually been effected and that I wasn't just being pranked by my so-called "friends". Especially if the space in my skull were filled with something that weighed about the same as my brain. (The name 'Airhead' might not be entirely accurate in such circumstances, but that's not uncommon with names with the form of descriptions.)

And on that note, another convenient benefit of the procedure would be some extra storage space. And not just for my jiujutsu tournament medals. If I stored my iPhone, address book, calculator, and related items in my cranial cavity, would I thereby de-extend my mind? Or would I be *extending* myself by reaching *inside* my skull? The mind boggles…

So the situation involves the spatial displacement of my brain and body. And I want, for now, to focus on the question of *where* I am. It will *seem* to me as though I am located wherever my body is located. My body is, after all, the nexus of all my perceptual input and behavioral output. So it couldn't help but to *seem* to me as though I am located there. (Ask yourself why you believe you are located where you are now. Is it because you can somehow sense where your brain is located? Or is it because the nexus of your perceptual intake and behavioral output is centered on your body?) But where am I *really*?

The brain-lubber will say that *I am where my brain is*. According to the brain-lubber, I am experiencing a sort of illusion. I *seem* to be displaced because my sensors and effectors have been displaced. The situation, the brain-lubber continues, is not unlike that of someone who undergoes a vague sense of displacement when remotely operating a robotic arm seen though a video camera. The illusion in the current case is just more comprehensive and more difficult – perhaps impossible – to shake. On the lubber's view most of my thoughts involving most demonstratives and indexicals would be strictly false. "I am now in front of my computer" would be false, though it would *seem* to be true. But there would be no problem with temporal indexicals: "I am now tired" is unproblematic.

The BBE-lubber will (I assume) reckon the mind as residing in two different locations, since the brain, body, and environment are now divided between these two locations. Some implications of this will be explored in subsequent sections.

At first blush, this case looks easy for the functionalist: since the *functions* haven't been changed, nothing about the mind should be changed. True, but functionalism doesn't usually concern itself with *where* questions. Even if a mind is nothing more than a functional system, where is this mind located? We are led to ask: *where are functional states*? One answer would be that functional states are abstract things and have no physical location. This response has nothing much to say about the location of

the mind in the scenario envisioned, except maybe that the mind, strictly speaking, has *no* location. I think this would be an unfair characterization of the functionalist position, though. A more charitable view would be that what we are concerned with are *instantiated functional states*, and so the functional states are located wherever the physical implementation of those functional states is.

If the functionalist appeals to physically implemented functional states, the possible answers will parallel the answers provided by the lubbers. This is because the brain was the implementor of the states before and after the procedure. The functional states are implemented in: (a) the brain only; or (b) the brain plus the body, especially the perceptual and behavioral (input/output) apparatus; or (c) all of that plus the relevant environmental objects perceived and acted upon. And the mind's location can be read off of this. It will be in the vat or distributed across the vat and body/environment.

What about contentualists? Regardless of whether we are talking about narratives, maps, models, or whatever, where is such a content-defined self located? No contentualist has directly addressed this question to my knowledge. Since the vehicles of the representations and the locations involved in contents typically coincide, the question hasn't been pressing. Brandom doesn't discuss it, nor does Ismael. Dennett is one of the few to point out that there is a significant question here, but he doesn't explicitly take a stand on an answer. Though it seems that he is a proponent of the view that the self is located where the (current) body is. Nevertheless, even if this is Dennett's position, I don't recall him ever providing an explicit argument for it. But it is worth trying to construct one, if for no other reason than that, as a contentualist myself, I believe it is the correct view. Here is a shot at it.

The reason the self is located where the body is is that this is what provides the *point of view*, and the point of view – the nexus of perceptual input and behavioral output – is what anchors and structures key locative elements of the relevant content-structure. The contentualist should treat subject location as defined by the various locative – primarily indexical and demonstrative – contents that are part of the content-structure. To a first approximation, to be refined shortly, if perceptual beliefs to the effect that I am between the Central Library and the Cognitive Science Building (because I can see them *there* and *there*) are part of the content-structure that defines me, then *that* is where I am. If the self is the center of narrative gravity, then it would seem to follow that when the narrative includes statements such as "I just left the

parking lot and am now between the Central Library and the Cognitive Science Building", that is where the self is. And these locative beliefs are keyed to the location of the body.

Some contentualists might resist this. But I think they would be wrong to do so. When a contentualist claims that the mind is a virtual entity like a center of narrative gravity or a model of some sort, then clearly what they have in mind is some manner of virtual or abstract thing *defined by the contents involved*. The word 'model', like 'experience', is ambiguous, and can refer either to the content or to the vehicle. But clearly such theorists aren't arguing that the self is the *vehicle*, as though if we found the cluster of neurons that were constructing the narrative or model, we could say "Ah! There is the self! Right there behind the thalamus!" No. Certainly not. The entity *described by* the narrative, or *modeled in* the model, is what is in question. And all features of these entities, including their locations, are determined by the relevant contents.

There are four points I wish to make about this before moving on.

First, the contentualist can easily embrace (and I think needs to embrace) the idea that the contents that play a role in self-location are *broad* contents. It is because I am in causal perceptual contact with *that* library and *that* building that my perceptual contents concern *them*, and accordingly, that I am located between *them* and not between their duplicates on Twin Earth. Rico will be located between the duplicate buildings on Twin Earth because his broad contents concern the objects in *his* environment. It is only by insisting that the contents are wide that a model or narrative constructed by my brain is a model/narrative of *me here*, as opposed to a model/narrative of *Rico on Twin Earth*.

Second, the locating contents are very fine-grained. More fine-grained than "I am in front of the library". Contents that course-grained would not distinguish different subjects standing roughly in front of the library. Rather, in play are very fine-grained contents including any and all perceptual contents required for perceptually guided action in the environment. When I am walking to the library, I need to know not just that the library is in front of me, but that *I am in danger of treading on that dog poop on library walk*, and that *I need to move a smidge to the right to avoid bumping into the side of the entrance*. The relevant demonstrative and indexical contents will easily

isolate one body as the anchor of the point of view. And note that they concern not just perceptual intake but behavioral output as well.

Third, on this view, subjects will typically be located where the body is, since it is the body, with its perceptual and behavioral capacities, that provides the experience upon which most relevant indexical and demonstrative locative beliefs are largely based. This is right, but it needs to be understood in the appropriate way. The danger is in being too cavalier about deciding what counts as the "body". A more accurate formulation would be that the self is located where the point of view is, and that the point of view of a self is defined by a nexus of perceptual inputs and behavioral outputs. Wherever this nexus is, there you will find the point of view, and so there you will find the self. And typically the biological body is where this nexus is. If, for some reason, something besides the biological body becomes the nexus of perceptual-behavioral efficacy, then that will be the "body" for locative purposes.[10]

The fourth and final point is a crucial refinement. The example of me being located between the Central Library and the Cognitive Science Building is open to the following *obvious objection*. Surely, if I am located in the Gaslamp District of downtown San Diego but am looking at two video monitors, one to my left projecting a live video feed of the Central Library and another to my right projecting a live video feed of the Cognitive Science Building, then I am not suddenly transported to the UCSD campus – even if I am momentarily unaware that I am looking at video monitors and think that I am directly seeing these buildings in my egocentric space.

Recall Gareth Evans's (1982) example of a person in a ship on the ocean surface whose sensory inputs are coming from a submarine located on the ocean floor and whose hands and feet are driving the submarine's propulsive apparatus and mechanical manipulators:

> One envisages, for example, a television screen showing pictures sent back from a remotely controlled submarine on the sea bed. Some straggly bits of seaweed appear, and so on. It seems that we can throw ourselves into the exploration: 'What have we here?', we say, or 'Here it's mucky.' (Evans 1982, p. 164)

[10] Notice that on this analysis there is a tighter dependence than might have been supposed of contentualist views (which can seem hyper-intellectualized) on considerations that would appeal to proponents of embedded/embodied approaches to understanding the mind (which definitely skew hypo-intellectualized). I will return to this in the final section.

> Let us elaborate the story, so that the submarine is equipped with limbs, excavators, etc., and a means of propulsion remotely controllable by the subject. And let us consider a highly trained subject, who can manipulate the limbs thousands of feet below him like the experienced driver of a mechanical excavator. By making this addition, we have certainly added to the strength of the subject's tendency to use 'here' and 'there'. (And equally 'this' and 'that': 'That's a remarkable fish.') The tendency would be especially strong if we insulated the subject from the sounds, smells, sights, and so on around him. (Evans 1982, p. 166)

It might seem that as I describe it, the contentualist position must maintain that the person would actually be on the ocean floor in this case, or actually be between the Central Library and the Cognitive Science Building. I think, however, that the contentualist can agree with Evans that in the case as described:

> The subject can *play* at being where the submarine is ('Here it's mucky'); he can *play* at having that mechanical contrivance for his body ('I'll pick up that rock'). But *really* he is (say) in the bowels of a ship on the surface of the water. This is not just one view he can adopt if he likes; *it is the view to which everything in his thinking points*. (Evans 1982, p. 166, emphasis added)

The reason is that it is not *just* the current perceptual contents that determine the locative aspects of the relevant content-structure, but the *entire structure*, the structure that includes current perceptual inputs and behavioral outputs but also beliefs concerning one's body (as the effector of actions, etc.), one's beliefs about one's worldly history, one's ability to locate oneself in the world and to navigate within it, and so forth. In Dennettian terms, if the bulk of my narrative has me as being located in my biological body in the surface ship, but then there are a few lines of narrative about how I am trying to move *that* rock from *here* on the ocean floor to *that spot over there*, then this isn't sufficient to change the location of the *center of narrative gravity* to the submarine. A few such perceptual beliefs don't override the mountains of other beliefs and other location-relevant contents that the subject has in this case, the bulk of which point to the biological body as the self's location. But Evans continues:

> Perhaps we can tell the story of the submarine in such a way that the subject's location in the surface vessel becomes less and less important to him. He does not move; he becomes insensitive to the sounds and smells around him. It might be possible (with enough of this sort of thing, and perhaps some surgical changes) for us to think of the submarine as *his body.* Then the centre of his world would be down on the sea bed, and his utterances of 'here' and 'this' could go direct to their objects without the need for conceptual supplementation. The precise details of this case do not matter; nor does its ultimate coherence. Let us grant that it is coherent; it does not affect the point I am trying to make. For now, of course, any information he received from the surface vessel – some sounds breaking through, say – would have to

> be thinkable from that vantage point (if incorporated into his thoughts at all). He would think *'Somewhere up there* someone is whispering', and his grasp of the thought would have a conceptual ingredient (involving the notion of *where my computing centre is).* (Evans 1982, p. 167)

Evans's point here is that the subject's location is woven into the narrative, into the content-structure, in a more nuanced and pervasive way than just being determined by whatever current perceptions indicate. For Evans, the important factors cluster around the subject's ability to locate himself in the world, to know which object he is in the world. Though the contentualist need not embrace precisely Evans's way of working locative contents into the content-structure as decisive, it is one reasonable approach. One can imagine, as Evans does, situations on either end of a spectrum. On one end would be situations in which the person has been operating the submarine only for a few minutes, her self-locating beliefs and skills (or center of narrative gravity, or whatever) all keyed to her body on the ship — she may have no idea where the submarine is and may not really care. On the other end of the spectrum would be cases in which a subject roams the ocean for years as a submarine, happily traveling between San Francisco to Hong Kong, and losing interest in, and perhaps even memory of, her brain and biological body — which, if she remembers correctly, were on a ship somewhere, perhaps they still are (*Who knows? Who cares? Playing with whales is just so much more fun than being a dirt-bound human...*). In such a case, "everything in [her] thinking" would now point to the location of the submarine as *her* location.

Exactly what sorts of transitions in the content-structures account for the difference between cases of the first sort and cases of the second is not the current topic. Evans, Dennett, Brandom, Ismael, and I will all have different answers to this. The important points are (a) that the contentualist position is that in cases of the second sort, the subject is located where the submarine is located, and (b) that the contentualist is not caught flat-footed by the obvious objection.

To recap before we move on: in this section I introduced the *Airhead* manipulation by way of a detailed version of the Dennett thought experiment and described how three major categories of theories of mind might reply to the *where* question raised by that experiment. Lubbers place Airhead's mind in the brain or distribute it between the brain and body; functionalists follow the lubbers because the brain/body/environment is what is implementing the relevant functional states; and the contentualists, insofar as they embrace the proposal put forward concerning how

they would locate subjects (which they should), place the mind at the point of view, *a.k.a.* the embedded body.

## 2. Temporal displacement of the basis: *Rip*

The brain and body are coupled systems: the brain gets signals *from* the body and sends signals *to* the body; the body gets signals *from* the brain and sends signals *to* the brain. Both of these information streams are continuous, real time, and mutually dependent. What signals the brain gets from the body depend, in part and in real time, on what signals the brain has sent to the body. If the brain just issued a command to close the eyelids, then this will impact the signal coming in along the optic nerve, to take a fairly banal example.

To emphasize how important this is, consider what happens when this real-time interaction is disrupted. Suppose that all signals from the brain to the body are delayed by, say, three seconds. This could be done by taking the information passing though the interfaces from the brain to the body, recording them to a hard drive, and then reading that same data stream three seconds later and passing it along to the body. This would be like taking one side of a telephone conversation and delaying it by three seconds. Such delays would make the brain's job very difficult. It might see an incoming football on a collision course and issue a command to duck, but the motor command would be delayed, and nothing immediately happens. The person would feel paralyzed and helpless as the football smacked her in the face, and then, three seconds too late, her hands would cover her face and the duck would be executed. Conversations would be difficult — even a delay of a half second or second make conversations challenging on cell phones, with the interlocutors often inadvertently speaking over each other. With a three-second delay in her vocal output, the person would find conversations very challenging. The fact that real-time feedback is crucial for just about any behavior means that such a delay would render anyone to whom this three-second-delay procedure had been done critically impaired. A delay of even a few hundred milliseconds can cause serious problems.

But what if, in addition to the interface from the brain to the body that is introducing a three-second delay, we arrange it so that the interface from the body to the brain *sends all signals back in time by three seconds*? Suppose, for example, there is a second hard drive that records the data stream that is headed to the brain. Then this

data is read off that hard drive and passed through a device that sends that data stream three seconds back in time and makes it available through the output port. In practice, this device would exhibit the behavior of being able to accurately predict (through its output port) whatever signal will be given to its input port in three seconds.

This would have the effect of canceling out the three-second delay from the brain to the body. So in the situation described above, if the incoming football were in front of the subject's eyes at exactly 12:00:00 pm, the signal from the retinas would be processed by the device described above and sent back to 11:59:57 — three seconds earlier — and then passed through the optic nerve. The brain would then, at 11:59:57 (plus the hundred or so milliseconds that normal processing takes) issue a duck and cover response. This response would be delayed by three seconds and would be sent to the body at 12:00:00 (plus the nominal processing time) — precisely in time to dodge the football.

Imagining myself into this situation is no more difficult than in the first manipulation. I would be getting apparent immediate feedback from my perceptual apparatus, immediately reflecting everything I am doing. When I close my eyelids, the scene immediately goes black. But a knowledgeable outside observer would see that my point of view is displaced three seconds into my brain's future. Of course, there is nothing sacrosanct about three seconds. The displacement could be ten seconds or a thousand years.[11]

Suppose we have a subject who undergoes this procedure and that the interface from their brain delays signals by, say, 1,000 years, before passing them to their body — which was frozen when their brain was removed and thawed out 1,000 years later so that it would be around to accept these delayed signals. The interface from the body to the brain sends the signals back in time by that same amount. We can call this subject *Rip*, in honor of the fable of Rip van Winkle.

And to put things into first person, suppose I am Rip. My predicament here will be more extreme than in the spatial displacement case. Not only might I be in the peculiar

[11] My colleague Craig Callender, who specializes in the physics and metaphysics of time assures me (personal communication) that this sort of situation, involving sending information back in time, is not a big deal. Apparently just bouncing the signal off a time-reversed galaxy will do the trick.

situation of seeing my own brain in a vat, but my brain may be long dead, floating in a jar of formaldehyde. It might be dissected. I could even be informed that it simply no longer exists, having been thrown in a compost bin when it stopped functioning. Imagine the indignation.

But it's not all bad. I would be, in a certain way at least, temporarily immortal — or at least immune to many normal causes of death. As with Airhead, shooting me in the head would be painful and inconvenient, but it wouldn't turn out my lights. But though Airhead's malefactors couldn't do him in *this* way, he'd be toast if they found his vat. *My* malefactors, on the other hand, are out of luck. Try as they might, there is nothing now in the universe they can do in order to turn out my lights. The best they can do is to eliminate my point of view by destroying my body and sending me into a sort of disembodied limbo.[12]

Clearly, the brain-lubber will claim that I am my brain and, like my brain, am residing in 2015, but my point of view has been temporally displaced. The *BBE*-lubber will say (I presume) that my mind is split between two times, along with my brain and body. There is really no reason for the lubbers to treat temporal displacement any differently than spatial displacement.

The functionalist will claim, as in the previous case, that I am either located (spatially and temporally) at my brain, or maybe distributed across my brain and body (and possibly environment). The possibilities for the various forms of functionalism discussed in Section 1 apply straight-forwardly to the case of temporal displacement.

The contentualist will maintain that the mind, or self, is in 3015 – *with no brain anywhere to be found*. In the same way that the demonstrative and indexical contents supported by experience in the Airhead case fix the spatial location of the content-defined entity that is the self, so too the *temporal* demonstratives and indexicals fix the temporal location of the self – in this case, 3015.

[12] In the example as constructed, the motor outputs that drive my body are coming from a hard drive that recorded these signals from my brain back in 2015, and all the sensory information being picked up by my body is being sent back in time to 2015. This means that another way my malefactors could remove my point of view would be to destroy either of these interfaces. It won't turn out my lights, but it would either paralyze me or anesthetize me (or both) depending on which interface gets compromised.

This is going to be an important and slippery issue, so I want to get a solid example on the table. Suppose there is a fireworks display on July 4, 3015, that I see and hear. I have perceptual experience and think about this experience in a way I would verbalize as "I saw the red fireball explode an instant ago, and in a moment I will hear it." These temporal contents codified by expressions such as "an instant ago," fix the temporal relations of my mental states. Assuming that I did have experience of the exploding red fireball, and that my prediction that I will hear the explosion in a moment bears out, my thought *I saw the red fireball explode an instant ago, and in a moment I will hear it* occurred temporally between my *seeing* of the fireball explosion and my *hearing* of it. This is a specification of the temporal relations between *mental states* (in this case, perceptual states and thoughts), a specification that the brain-lubber and contentualist will agree on. It's just that the lubber thinks that this entire series of mental states is manifesting in 2015, while the contentualist takes it to be manifesting in 3015.

According to the contentualist, but not the brain-lubber, my visual experience of the explosion, the thought *I saw the red fireball explode an instant ago, and in a moment I will hear it*, and the auditory experience of the explosion all occurred *after* the actual explosion (a specification of temporal relations between mental states and a *physical event*), which occurred in 3015. This is what it means for the temporal indexicals, anchored to external events, to determine the temporal features of the content structure that defines me. It is the temporal analogue of what was said in the previous section concerning spatial location being determined by spatial demonstratives and indexicals.

Before closing this section, I want to describe a feature of the situation that will highlight another radical consequence of the contentualist position. Suppose that I am in the situation described in this section and take myself (perhaps truly, depending on who you ask) to be in 3015. I decide to go into a dark room to just be alone with my thoughts for a few minutes. I turn out the lights, lie down, and silently think through my plans to go to the fireworks show later that evening. I am not moving, and there is little or no sensory input. The hard drive that feeds motor signals to my body will be largely quiescent, as will the interface from that drive to my body. And the interface that is sending signals back in time to my body is also quiescent. So my several minutes of internal monologue – my train of thought about who to invite, whether to drive or take Uber, where to go for the best view and so forth – is all happening not only without any brain currently on the scene, but with no activity to speak of in the

input or output mechanisms. The contentualist's position will be that despite nothing physical happening at the time that is in any way connected to brain functioning or input or output, there is nevertheless a train of thought progressing there and then. There are only a brainless body breathing and digesting, and some quiescent interfaces that would become active if any sensory signals came in or if I decided to move.

To many, all of this will seem to constitute a *reductio* of the contentualist position. And I imagine that some contentualists may be experiencing pangs of buyer's remorse. A subject can exist, think, and behave normally, with no brain (or any other information-processing infrastructure or material substratum) currently around? While not a *reductio*, it certainly is a bitter pill. But lest the contentualist's opponents start feeling too good about themselves over this, let me point out that there's a bitter pill buffet on the way.

## 3. Spatiotemporal dispersal: *Scatterbrain*

Now I wish to add a layer of complexity to the previous manipulations. Maybe one and a half layers. The first layer is this. In both of the previous manipulations, we treated the brain and body as two separable entities in close reciprocal causal interaction. Which of course they are – though 'close' must be understood in causal terms, not spatiotemporal terms. And we have been imagining the consequences of technology that would allow for those causal connections to be stretched across space and time while maintaining the same causal structure. There is nothing, however, that requires us to keep the brain anatomically intact. The Airhead manipulation might have been pitched as one in which only one cerebral hemisphere was removed, and the other kept in the body, with all the causal interactions between the removed hemisphere and the body+non-removed hemisphere kept intact. Or both hemispheres might have been removed separately, so long as all the causal interactions between the three components — (a) left hemisphere, (b) right hemisphere, and (c) body — were preserved. Indeed, the brain could be divided into any number of chunks (right frontal lobe, thalamus, left primary visual cortex, and so forth), so long as all the causal influences are preserved by appropriate interfaces. By *MPH*, the mind would remain unaffected.

So let's suppose that the subject's brain has been divided into, say, 10 parts that have been spatially dispersed throughout a vast region, but with all causal interactions maintained through interfaces as before. I'll call this subject *Scatterbrain*. One consequence is that there won't be any tidy location for the brain-lubber to think of Scatterbrain's mind or self as inhabiting. In fact, there isn't really any tidy location where any of his thoughts or feelings would be located on the lubberist take on things. It is plausible that any given conscious perceptual state is supported by several brain areas: perhaps visual cortices, some thalamic areas, and maybe some frontal and temporal lobe bits as well. If these parts are spatially separated, possibly by kilometers, where would that experience be *located*, according to the brain-lubber? And where would the mind be that has such experiences? There's no easy answer here. It would seem that the brain-lubber, indeed all the lubbers, must say that Scatterbrain's thoughts and experiences are dispersed, perhaps over a vast region.

And what about indexical thoughts, like "I am here"? In the case of Airhead, the brain-lubber supposed that 'here' would refer to the location of the brain. But what is its referent for Scatterbrain? Again, there is no easy answer. Is the referent the set of some number of small discontinuous locations where the brain parts are? 'Here' is semantically quite flexible, but if anything inflicts fatal violence on the notion of *here*, it is a set of widely distributed discontinuous regions. Perhaps it refers to a large volume, something like the convex hull of those parts. If those parts are spread wide, this could be quite a volume. Will this volume change depending on the specific parts used in the thought? If not, why not? If so, can the referent of Scatterbrain's *here*-thoughts be altered significantly by taking one small little-used part of his brain and flinging it over to Mars? Or perhaps in this situation his *here*-thoughts lack a referent. His thought that "It is warm here" when he is in (or takes himself to be in) a sauna would be false at best (on some views, possibly meaningless). Perhaps Scatterbrain simply has no *here*. None of these options is available on the cheap. The brain-lubber's convenient retreat has become markedly less convenient. This is similar to the situation the BBE-lubber was in in the spatial displacement case, since even there the physical stuff that was important for the BBE-lubber was dispersed. We have just landed the brain-lubber in the same boat.

The functionalists' responses will be parallel to those of the corresponding lubbers. The contentualist is in no more difficult a situation than before. The contentualist

doesn't appeal at all to where the information-processing machinery is, and so its dispersal is irrelevant.

Now for that extra half-layer of complexity. Scatterbrain's brain parts might be *temporally* as well as spatially dispersed, with some operating now, others in the distant past, and yet others in the future. The implementation requires no new methods beyond those already in play, it is just a more complicated set-up in order to maintain all of the reciprocal causal interactions.

Suppose that this is my situation, suppose that I am Scatterbrain. What should I make of the possibility that my brain is spatio*temporally* dispersed? That my brain has no location that establishes a referent for my *here*s is one thing. But that I have no *now* is something altogether more disruptive. I sit here (?) looking at a computer screen, its visual appearance, the feel of my fingers on the keys, all working in harmony with my "internal" stream of thought about what sentences will best express these ideas I am trying to articulate. And there seems to be the usual temporal coherence: I am formulating my ideas now in part in response to what I was thinking about a moment ago; my train of thought was interrupted by an email that just came in. There certainly *appears* to be a determinate sequence of experiences and thoughts that are embedded in, and synced to, both their own well-defined internal milieu, as well as to the evolving environment around my body. How can *this* be taken away without removing everything deserving of the name *self*? What would it even mean for it to be an illusion?

To put this in bold relief, recall the fireworks display. According to the brain-lubber (and the BBE-lubber and all versions of functionalists), to Rip it only *seemed* like his thought *The red fireball exploded an instant ago, and in a moment I shall hear it* occurred between the actual explosion and the arrival of the acoustic waves. But for Scatterbrain, the situation is worse – much worse – for it also only *seems* to him as though it is a part of a temporally coherent stream of thought at all, one that goes from *Behold that (visible) explosion!* to *The explosion was just an instant ago, and in a moment I shall hear it* and then to *Sure enough, there's the sound of the explosion!* In the case of Scatterbrain, if we try to pin down the temporality of these mental states by reference to the neural processing underlying them, we shall come up if not empty handed, then at best with hands full of jumbles and knots. Even if the individual mental states (*this* perception, *that* thought) had a definite time of occurrence because they were entirely housed in one of the brain chunks, they might easily be out of order:

my *seeing of the explosion* might have occurred long after my thought that it had just occurred, if the visual cortices are operating in 2715 but the auditory cortices are operating in 2215 (or whatever the details). But even the individual mental states will most likely lack any defined temporal position for the same reasons they would lack any definite spatial location. The neural hardware supporting each mental state might be temporally dispersed.

The contentualist is faced with no new challenges here. The brain is merely supplying the information-processing infrastructure. Its location and temporality are irrelevant so long as its causal structure supports this role, and the causal interactions between all the relevant parts are being maintained. What counts are the contents grasped as described in Section 1. My *thought* about the explosion occurs between the physical event of the explosion and the arrival of the acoustic waves because that is where the contents define it as being, since those contents are causally anchored to the relevant external events. The spatial and temporal location of the neural machinery whose computational prowess constitutes/constructs the vehicles of those contents is beside the point.

The idea that the mind implemented in the physical stuff of the brain might (a) lack any defined temporality, supporting no actual *sequence* of thoughts or any *sequence* of mental states, while (b) still being the mind of a normally functioning person, should give anyone pause. But there is a move open to the lubbers and functionalists to try to avoid this dire conclusion. While it is true that the brain parts are temporally dispersed and scrambled, it is also true that in order for each of them to fill its causal role correctly with respect to all the other parts and the body, it must be synced with the others in what we might call *processing time*.

First an analogy: a musical group can create a recording of a song in two ways. First, all the musicians can simultaneously be in the recording studio and perform the song together, with this group performance recorded. Second, the musicians can record their parts independently at different times, with these individual performance recordings synced to create a recording of the song. In the latter kind of case, even though each musician performs their part at a different time, there is still a time frame that each must sync to. If the song is four minutes long, then there will be an abstract four-minute time frame such that each musician begins at its beginning and ends at its end. There is an abstract temporal framework here, with an order and metric. And we can describe events in a *well-defined though abstract* temporal structure clearly and precisely: *When was the first drum fill? Four seconds (or beats, or measures) before the*

*accordion solo. Sadly the vocalist came in late after the bridge*. (Notice that the vocalist can be *late* in the abstract temporal frame even if the vocals were recorded days or years *before* the other instruments in real time.)[13]

Analogously, we can see that even if Scatterbrain's brain is temporally dispersed, there must still be an abstract *processing time* that each part must sync to. And this includes the environment-embedded body, of course. This then is the lubbers' and functionalists' countermeasure to the suggestion that on their view, Scatterbrain would have no definite mental timeline. There is a well-defined temporality, in terms of this abstract processing time. If Scatterbrain survives for twenty years in this brain-dispersed state, then there will be an abstract twenty-year span that his body and each brain part are all synced to. Each part will live and operate for twenty years. If the first thing that happens when Scatterbrain wakes up from the procedure is that he hears a question and answers it over the course of five seconds, then the first five seconds of each of the parts' operation, regardless of when it is doing that processing in real time, will be geared to managing that question and response.

I think there is something deeply correct about this. But note that an *abstract* temporal interval *as such* is not identical to or even coincident with any part of the temporal history of the actual universe. It *might* be instantiated, implemented, or anchored in an actual temporal interval. Accordingly, questions can be raised with respect to the mental timeline putatively rescued by this gambit. Is this abstract mental timeline coincident with (or anchored in, or instantiated in) any actual temporal interval? And if so, what interval? If the mental timeline is not coincident with any actual temporal interval – that is, if it remains purely abstract – then it seems that the mental life in question would be merely abstract.

And notice that we have, *ex hypothesi*, maintained all causal interactions despite the brain parts' temporal scattering. This would mean that BPH is satisfied.

[13] This abstract time frame can be implemented in different real temporal intervals (e.g. the vocalist can do her four-minute part on Monday afternoon from 3:28pm to 3:32pm, and the pianist can do hers on Friday morning from 9:05am to 9:09am), but it can also be transformed in various ways. The guitarist might, when recording his part, slow the tape with the other recordings on it down to 50% of its normal speed and record his part over eight minutes. When played back in four minutes, the guitar parts will be an octave higher and with more apparent virtuoso speed than the actual performance that was recorded. The transformations might also be non-affine. I won't explore these possibilities, as they make no essential difference to any points I wish to make.

Regardless of what you think about the mind, if any, supported by the scattered brain, it would seem difficult to deny that overt behavior could be maintained. But a situation in which BPH holds but MPH does not is precisely a situation in which we have a "philosophical zombie." Anyone who wishes to deny that Scatterbrain has a genuine mind would need to either accept the idea of philosophical zombies or somehow try to deny that BPH would hold in this situation. But the situation is specifically designed to be one in which any physicalist would be all but required to accept BPH (see my remarks in Section 1 on that topic).

I'm really not sure what to make of the idea that the temporality of my current train of thought – or that of any normally behaving subject to whom we wish to credit a mind – is not coincident with any interval of actual time in the universe, or that it is *merely* abstract. This doesn't seem like a promising option to pursue.

If, on the other hand, the mental timeline *is* coincident with some actual interval of real time, then which interval? The only sensible option would seem to be to make it coincident with the interval over which one of the interconnected components operates. And the options are either one of the brain parts or the embedded point of view (i.e. body). The proposal that it be some preferred brain part – the thalamus, the pineal gland, or whatever – faces the obvious challenge that this part could easily be divided with its parts temporally dispersed. There is no defensible refuge to be found down that trail.

This leaves the embedded point of view. This option is attractive for independent reasons. One is the fact that this component is the only one that can't be divided without disrupting the overall system's functioning. If I split the *body* (or whatever it is that is the business end of perceptual input and behavioral output) into two halves and temporally disperse them while leaving their causal connections intact, the mere fact that each part is in a different environment now will render normal mental and cognitive operations highly disrupted. (If you are offered a choice between becoming Scatterbrain or Scatterbody, you should definitely opt for Scatterbrain.) Another related reason is that this is the interval where the mental rubber meets the worldly road.

But whatever the reasons – whether the unattractiveness of all other options or the positive appeal of this option – the embedded bodily point of view seems to emerge as the only viable candidate for the component that defines the actual temporal interval in which the mind's temporality is anchored. And of course, this is exactly as the contentualist would have it. For it is the embedded point of view that is the source of the relevant indexical and demonstrative contents. The lubbers' and functionalists' attempt to save a mental timeline by appealing to processing time may not flat out land them in the contentualists' camp, but it points them in that direction and provides an insistent shove.

The manipulations discussed in Sections 1, 2, and 3 were manipulations of the spatial and temporal location of the brain and its parts and the embedded body, together with an assessment of how this would affect the location of the mind, according to various views of the relation between the brain (/body/environment) and mind. And everyone ends up taking some lumps. The brain lubber has an easy time with Airhead and Rip, insisting that the mind is where the brain is. The BBE-lubber is in slightly less comfortable water, since the mind will be split up. For all lubbers the water goes from uncomfortable to hazardously hot with Scatterbrain. The functionalists have the same issues as the lubbers. The contentualists' biggest discomfiture is the consequence that a fully minded and normally functioning person might exist after their brain is long gone.

## 4. Synchronic Individuation: *Raid* and *Janus*

The next three manipulations will explore a very different sort of problem space. While the scenarios described in the first three manipulations hinged on the contributions of brain and body (embedded point of view) to the *location* of the subject in space and time, the next three manipulations will explore the contributions to *synchronic individuation* and *diachronic identity*.

Let's return to the original spatial displacement scenario with a brain removed from the body and left intact, as with Airhead. But rather than one brain, there are two qualitatively identical ones. How this might happen is unimportant, but for those who like their thought experiments complete with an origin story, we can suppose that I was the result of a cloning process that produced two identical versions of my

brain+body and that the displacement procedure was performed on both clones immediately upon their artificial birth. However, rather than each brain being connected to its own proprietary body, the two brains *unknowingly* are jointly connected to a single body. Really, whether one of the brains is still in the body is irrelevant, but let's keep things symmetric by supposing they have both been removed. All sensory signals from the one body are duplicated and sent to each of the two brains so that each gets precisely the same sensory inputs as the other. And the brains' motor outputs to the body are combined by an AND gate, so that any given motor-command signal is passed along to the body if and only if that same signal is produced by both brains.[14, 15]

The situation is analogous to a common method of making hard drives more robust. A RAID(1) array is a set of hard drives that are, essentially, mirrored. If you have two 1 TB hard drives, you could use them both such that you have a total of 2 TB of storage available to you. This would be a RAID(0). Or you could set them up in a RAID(1) array, and in such a case the drives would be essentially mirrored to create a single virtual 1 TB drive that is hardware-redundant. The cost of the RAID(1) is that you end up with only 1 TB of storage rather than the 2 TB you get with a RAID(0). But the advantage is that the virtual 1 TB drive is more robust to hardware problems. If one of the drives malfunctions, the virtual drive survives. With this in mind, let's call the subject in this situation *Raid*.

Like the situations described in the prior manipulations, we would expect from MPH that from the first-person point of view everything would seem to be precisely as it would have seemed had I not been involved in any outlandish experiments. I would,

---

[14] Since, *ex hypothesi,* both brains are identical and so will always produce the same commands so long as they receive the same inputs, an OR gate, or averaging the motor commands, or even having only one of them in control (the other one operating under an illusion of control since all of its outputs have no consequence), would yield the same behavioral result as the AND gate. But the AND gate has the theoretical convenience of making each brain *responsible* for the actions taken at least in the sense that if counterfactually, that brain had not undertaken that action (i.e. had not issued that motor command), the action would not have been taken.

[15] The original Dennett situation in "Where am I?" (Dennett 1981) had something akin to this going on. At one point in the narrative, Dennett learned that a computer program that replicated his brain's input/output structure was constructed and was run in quasi-parallel to his actual brain. Specifically, the brain was getting inputs from, and sending outputs to, the body. The computer program was also getting copies of those inputs, but its outputs were, unbeknownst to it, not sent to the body. Since its outputs were, *ex hypothesi*, the same as those of the actual brain, it did not notice that it was ineffectual. When he learns of this, the decision is made to allow the two control centers (brain and computer) to be swapped in as the one actually in control, while the other gets swapped out to an ineffectual illusion of control. The device that allows for this switching is not labeled and so it is not known at any time which – brain or computer – is in control.

of course, have no inkling of the brain duplication unless I was informed of the details of the procedure.

But first-person phenomenology aside, there is an oddity here. *How many* subjects with first-person phenomenologies are there in this circumstance? Let's start with the possibility that most naturally suggests itself, especially to anyone with brain-lubber-leaning intuitions: there are two subjects, one supported by each brain. I will suggest in a moment that it is not obvious that this must be the brain-lubber's position. But for now I will just remark that if I were in this situation and informed of the details, this possibility would be somewhat vertigo-inducing. What am I to make of the suggestion that there exists (somewhere? here?) a mind that is *exactly* like mine, that has every thought, every feeling, and every emotion I have, precisely when I have it? Keep in mind that these mental states are the same in a broad-content object-involving sense. It is enjoying *this very cup of hot chocolate* to precisely the same extent I now am, looking through *these very eyes*, moving *this very hand* to *this very delicious pastry* that I (we?) will be enjoying, and typing on this very keyboard that is now under my (our?) fingertips. I might *deduce* that this shadow-self is as puzzled about me as I am about it. And yet for all its unsettling closeness, I can't access it at all. Even other people's minds I can access to *some* extent. I can ask questions and get clearly recognizable responses that even in the worst-case at least give me some evidence that there is a mind behind the answer. But what question can I ask my shadow? Every question I contemplate (or vocalize out loud) appears to be acknowledged only by me. And there is never an answer, except the one I produce myself. I might infer that this is the answer my shadow is providing in its own internal monologue, as an answer to a hypothesized interlocutor that it assumes is asking a question. We can glimpse here a singularity in the problem of other minds: intimacy has become infinite, and accessibility is undefined.

While the brain-lubber *could* claim that there are two subjects, because there are two brains, I don't think this position is forced upon her. She might also say that there is one subject with one architecturally redundant brain. Why might this be acceptable? In the normal case, brains have a good deal of redundancy, and we don't normally take this to be a sign of multiple subjecthood. First, temporary or permanent disconnection or anesthetization of one hemisphere typically leaves a functioning subject (in, e.g., as in the Wada procedure, or in so-called split-brain patients). It is not obviously wrong to claim that the subject present in such cases is the same one that had more robust

and redundant causal hardware supporting it when both hemispheres were awake and running normally. Second, as things stand in normal brains, there is a great deal of redundancy at the circuit level. This is just good design, since if one component fails, overall system function doesn't collapse. This means that we need to make room in our thinking for subjects that have *at least some* redundant hardware whether we like it or not. The procedure I'm describing with two brains and one body *could* be seen, if a brain-lubber were so inclined, as just adding more redundancy than is usually the case. Though of course I suspect that most brain-lubbers will stick with the one-mind-per-brain response to this manipulation.

There is an analogous individuation problem in normal RAID(1) arrays. Does such an array produce two identical drives, or one virtual drive with redundant hardware? It is hard to imagine a scenario where an answer to this question in the context of hard drive data storage would be pressing. But it is perhaps not overly difficult to imagine such a context in the case of brains and persons. If Raid kills herself, was there one suicide, or were there two? If it was homicide, would the perp be facing one count of murder or two counts?

What about the BBE-lubber? This is a tough one. If in normal situations a mind results from an embodied brain, perhaps embedded in an environment, what happens when there are two brains and one embedding body/environment? If a knife is composed of one handle and one blade, how many knives are there when you have one handle and two blades (as can be the case with, e.g., a Swiss Army knife)? If the answer is either one or two, then there is a recognition of an asymmetry in the importance of one of the components, such that one of them is providing the individuation conditions that allow us to count knives. But if the answer is neither *one* nor *two*, then what? Can there be one and a half knives? Or an indeterminate number? (Perhaps an imaginary number?) Counting one and a half knives when one has two handles and one blade is perfectly reasonable. And saying that a Swiss Army knife with two blades is still one knife is also reasonable – but betrays an asymmetry in the importance of the handle vs. blade for purposes of knife individuation. But what would it mean to say of a situation like this that there are one and a half minds?

I'm really not sure what the functionalist would say, though it would seem that there is reason for her to claim that there is one mind here, because there is one functional organization. Recall Nelson's (1975) example of a functional system, a soda machine. It is defined by the following state table:

| | $S_1$ | $S_2$ |
|---|---|---|
| Nickel Input | Emit no output<br>Go to $S_2$ | Emit a Coke<br>Go to $S_1$ |
| Dime Input | Emit a Coke<br>Stay in $S_1$ | Emit a Coke & a nickel<br>Go to $S_1$ |

The inputs and outputs are not abstract: the insertion of nickels and dimes into metal slots and the outputting of cans of soda from dispensing bins are concrete if anything is. But what appeals to functionalists is the fact that the functional states, in this case $S_1$ and $S_2$, are independent of any particular material instantiation. The relevant functional states can be implemented in *whatever* physical system it is whose behavior is describable by this state table – mechanical gears, silicon chips, whatever. This suggests, perhaps even entails, that even in cases of redundant physical hardware there is still just one functional system. A soda machine manufacturer might implement the state table in not one but two programmable chips running in parallel, just to make the whole system more robust to physical damage. If one of the chips fails, the vending machine owner still accumulates coins and the consumers still get Cokes.

What *seems* to be doing the individuating of *identical functional systems* with the same state table are the concrete input and output interfaces. Suppose I have two soda machines, one in the lounge and another in the mail room, such that the state tables are implemented in numerically distinct but qualitatively identical chips. Even if, by some massive coincidence, they are always in the same state – because whenever someone buys a soda from one, another person buys the same soda from the other, with the same sequence of coins, at the same time – there are still two distinct functional systems, because the concrete input and output interfaces are distinct. *Their* identity is not in question. It is determined not functionally but by the usual mundane criteria that apply to coin slots and dispensing bins.

The reason there is (arguably) *one* soda machine in the case of a single machine with redundant state-table implementation, rather than two, is because there is *one* coin slot and *one* dispensing bin. At any rate, this seems to be a reasonable functionalist position.

Against those who resist this suggestion, it can be pointed out that for functional entities, redundancy is typically not taken to be an indicator of multiple identical functional systems. A paperweight is a thing that, regardless of physical instantiation, fills the function of keeping papers in place. If I have a 1 kg iron paperweight, it seems that I have *exactly one* paperweight, even if I could saw it in half and each half could serve the function just as well. The fact that the physically redundant functional system *could* easily be divided into two functional systems of the same sort is not typically taken to be sufficient reason to think of the initial redundant system as two separate functional systems. And even more tellingly, if I *started* with two iron paperweights and then, as part of a desktop organization program, welded them together, then I would have one paperweight: a redundant one made of materials that previously comprised two paperweights.

In the case of the two brains running under the hood of a single body, there is one concrete input/output mechanism, the body. It is not obvious how a functionalist could motivate the suggestion that the two brains are actually two separate implementations of a functional system as opposed to one hardware-redundant implementation of one functional system. One way (I'll discuss another later on) would seem to be by some sort of appeal to the underlying material substrate. I have just provided some barriers that anyone attracted to this tactic would need to overcome. But aside from that, the *whole point* of functionalism was to loosen the ties between mentality and its material implementation. It would seem that the functionalist can maintain numerical distinctness of the functional structures here only by making a significant concession to their physicalist (/lubberist) opponents, namely, by allowing the physical stuff to play a role in the individuation of the relevant functional systems over and above the role played by the relevant functional specifications.

There are other possible things a functionalist might say. I will let interested functionalists speak for themselves. But for now, I will use the expression *pure functionalist* for the functionalist who, despite physically redundant implementation, individuates functional systems on the basis of their concrete input and output interfaces. The individuation at issue is synchronic – diachronic identity will be addressed in the next section. *Dirty functionalists* are those who allow *physical*

*implementation* of the functional systems, in addition to the input and output interfaces, to contribute to their individuation. So in the cases described above, the pure functionalist will say that there is one soda machine (with a redundant controller chip), one paperweight (that could be divided into two), and one subject (with redundant neural hardware), while the dirty functionalist will count two soda machines (that share a coin slot and dispensing bin), two paperweights (in a single cube of iron), and two subjects (that share a body).

There are pure and dirty versions of the contentualist position as well. Contentualism as I have so far been describing it has been pure: the material substrate has been irrelevant not only to the location of the self but also its identity. Identity of subject has turned entirely on identity of content-structure, which has turned on the contents involved and their structuring relations. So according to the pure contentualist, since both brains are maintaining structures with the same wide contents (because they both are embedded in the same environment with the same body), and since both brains are structuring them in the same way (e.g. by writing Dennettian narratives or constructing Ismaelian models that don't differ in any details), then they are both redundantly maintaining the *numerically* same content structure and hence are redundantly supporting the *numerically* same subject. There is only one Raid, not two.

A *dirty contentualist*, like the dirty functionalist, allows the material substrate to play an individuating role in content structures. And so this theorist would allow that in this case, because there are two different brains involved in maintaining content structures, there are two qualitatively identical but numerically distinct content structures and hence two subjects, two Raids sharing a body. So for both kinds of dirty theorists, having distinct material substrates is *sufficient* to count distinct minds. The pure theorists will maintain that in these cases there is a single functional system or content structure with redundant hardware. In other words, distinct material substrates are not sufficient to count distinct minds. Rather, in such cases, there is just hardware redundancy.

What about necessity? That is, is a difference in hardware *necessary* for a difference in functional system (or content-structure)? To explore this, let's change the example. Suppose again two identical cloned versions of me, but rather than this being reduced to two brains and one body, it gets reduced to one brain and two bodies. Imagine that the qualitatively identical bodies are placed in qualitatively identical but numerically distinct environments, such as Earth and Twin Earth. The single brain is conveniently

located half-way between Earth and Twin Earth, at (where else?) Middle Earth. We should imagine all of the outputs from this brain multiplexed into two identical copies and sent to the two bodies, and all of the inputs to the brain formed by combining both bodies' signals. Again, for now, assume that no differences emerge in the evolution of events in the two bodies or environments. Divergences will be addressed later on. Let's call the subject in such a situation Janus.

What we are interested in is whether we have one or two subjects in this case: one subject, supported by one brain, embedded in two qualitatively identical but numerically distinct bodies/environments; or two subjects, one per body/environment, who happen to share a single brain. The brain-lubber's position is clear: Janus is one subject who is living in two distinct but experientially superimposed worlds. Everything Janus experiences is a superimposition of two different environments, and everything she does is executed doubly. The BBE-lubbers are faced with a different version of the same difficult counting question discussed above. It would seem that if there is an answer to the "how many minds?" question it should be a whole integer, especially when all parties are acting overtly like normal humans: "One brain embedded in one body = one mind; so one brain embedded in two bodies = 1.5 minds" seems less defensible than either *one* or *two* as a response. But if the answer is *one*, then these lubbers owe us an explanation for why they aren't simply brain-lubbers; and if it is *two*, then they have the same bullet to bite as the contentualist to the effect that a single brain can play a role in supporting two distinct minds.[16]

The dirty functionalist, like the brain-lubber, may very well say that Janus is a single subject embedded in two environments. Despite the fact that the input and output interfaces are doubled, since the functional system is implemented in only one substrate, there is only one functional system. Whether this is the position depends on whether the dirty functionalist takes it that a difference of material instantiation of functional system is not only sufficient for discerning a difference of functional implementation, but is also *necessary*. I shan't pursue this, but I believe that a dirty

[16] And not in the relatively uninteresting way that this might occur with a split-brain patient. In those cases, there are separate anatomical entities underwriting each mind/self (to whatever extent there are separate minds/selves). But in the case currently under consideration, it is the very same hardware involved in both.

functionalist who wished to resist the necessity claim would not lack assets. Some of which I'll develop now.

I think it is open to the pure functionalist (and maybe even to the dirty functionalist) to characterize this as a situation in which one physical device, the brain, is implementing two instances of a functional system. What is doing the individuating, on this view of matters, is not the physical implementation of the functional system (it is the *pure* functionalist's views we are currently discussing, after all) but the two distinct input/output interfaces. Keep in mind that on a long-arm or wide-content view, there would be *different* inputs supplied by the two points of view and hence presumably different functional states residing between inputs and outputs. For example, the subject on Twin Earth would, in suitable circumstances, be in a functionally-defined XYZ-desiring state, whereas the subject on Earth would be in a functionally-defined $H_2O$-desiring state, and so forth. If there is only one subject, it would seem that we must forego wide contents as normally understood. I will return to this issue in a moment when discussing contentualists.

As soon as we leave the purely abstract state-table specification and look at actual instantiated functional states, there would seem to be at least some *prima facie* reason to recognize the possibility of distinct functional systems implemented in the same neural hardware. For starters, the pure functionalist will point out, the same material object is quite commonly allowed to instantiate more than one functional system. Your paperweight might also be your alarm clock. And your alarm clock might also be your partner's alarm clock. And it need not be the case that you and your partner have to get up at the same time to share this alarm clock. If your shared bed straddles a time zone boundary, the same physical device might simultaneously serve the functions of (a) an alarm clock waking you up at 8 am (PST) to get up in time for your 9 am (PST) meeting, and (b) an alarm clock waking your partner at 9 am (MST) for his 10 am (MST) meeting. The same physical entity in this case supports two different specifications of its inputs, outputs, and interposed functional states. If the dirty functionalist wants to insist that the same material substrate cannot be an implementation of two separate functional systems simultaneously, then her work is cut out for her.

What about the contentualist? Addressing what the contentualist will say will involve discussing the individuation of *contents* – because that is one of the considerations involved in individuating content-structures – so let's take a detour

through an analogy that should prove helpful. The old-school method of producing a photograph involves a causal process in which light from an apple (say) impacts negative film in a camera, and after being developed, this negative is used to get the correct pattern of light to project onto photographic film, resulting in a photograph. Let's agree that because the camera was in front of apple A when the negative was exposed, the resulting photograph is a photograph of apple A and not of qualitatively identical apple B on Twin Earth.

But what are we to say of a photograph produced in the following way? A camera in front of apple A produces a negative, while a different camera in front of apple B on Twin Earth (or even regular Earth, for that matter) also produces a negative. The two negatives are qualitatively identical. Both are delivered to a single print shop on Middle Earth, and a single print is made by projecting superimposed images from both negatives simultaneously onto one piece of photographic paper. In terms of the specific arrangement of colors, the photograph is qualitatively identical to what would have been produced by either negative alone. Which apple, if either, is the resulting print a photograph of? It seems there are three (at least) possible responses one could give:

1. The print is not a photograph of either apple. The causal mechanism, because it does not originate from a single object, is an aberration. At least insofar as its status as a representation goes.

2. It is *one* photograph *of both apples*. That is, it is one photograph that carries a single content, and that content is *the two apples*. It is on this analysis similar to a photograph, produced in the usual way, of two apples side by side. Or perhaps better, of two translucent panes of glass, one behind the other. On this view one could produce a photograph *of a completed jigsaw puzzle* in the following way: place one jigsaw puzzle piece on a black table in the position and orientation it would be in if it were part of the complete puzzle, and take a photo of it (i.e., expose a negative); then remove that puzzle piece and place the next one down at the location and orientation *it* would be in if the puzzle were complete, expose a different negative, remove that piece, repeat the procedure for all pieces, and then sequentially project images from each of those negatives onto a single piece of photographic paper. If carried out correctly, the result should look like a completed jigsaw puzzle.

But I think it would be wrong to maintain that this *is* a photograph *of* a completed jigsaw puzzle. For starters, there isn't and never was a completed jigsaw puzzle there for it to be a photograph of. Saying that it would *depict* a completed jigsaw puzzle

might seem like an acceptable way to describe it, since it was, after all, contrived to give the appearance of being a picture of a completed jigsaw puzzle. But we shouldn't get confused between (a) a photograph that *depicts* X, and (b) a photograph *of* X. I expect that anyone who claims to resist that distinction might quickly flip-flop and vehemently *defend* the distinction if a doctored photograph (contrived via photoshop, or clever multiple exposures, or what-have-you) *depicting* them stealing a briefcase full of cash from a Mafia boss's office safe were provided to said Mafia boss. Whether one thinks that the distinction is best marked terminologically by the expressions "photograph of x" vs. "photograph depicting x" isn't the important issue. The important issue is that there is a distinction, and the first sort of case has a kind of content not carried by the second kind of case. This is all by way of trying to put pressure on the suggestion that the content of the photograph would be *of the two apples*.

3. It is really *two distinct photographs* implemented on one sheet of photographic paper. One is a photograph of apple A, and the other a photograph of apple B. One motivation for describing the situation this way is that it seems that this is precisely what one would say if there were a single piece of photographic paper that was exposed to two completely different negatives – for instance, one of an apple on a table and one of a sunny mountainside. The two images would be superimposed, but it seems to me that the correct thing to say is that it is two separate photographs on one piece of paper – though of course it is possible that intuitions might differ on this. But if that *is* the correct thing to say when the two negatives are qualitatively quite different, then it should also be the correct thing to say when the two negatives are qualitatively identical.

With a discussion of the photograph case in hand, let's return now to the contentualist view, according to which the individuation of content-structures hangs in large part on the individuation of contents. We can assume that in the Janus case, since there is one information-processing system doing the structuring, we are dealing with only one structure-type.[17] This would be like two narratives that are alike in all details but which refer to different things: a narrative written on Earth and a corresponding one written on Twin Earth would have the same *structure-type*. The

[17] Actually, depending on how one individuates them, there might be two. But even if there are two, they process information in precisely the same way, which is the important thing for this point.

question, then, is how many sets of contents are being structured in Janus? Analogues of 1, 2, and 3 above are *prima facie* available to the contentualist to get a handle on this.

1. There is no subject at all in this case, because there isn't really a content-structure. Every putative content is an aberration owing to its causal history. Because of the lack of a unique causal embedding, there are no *wide* contents. On this account, while there would not be a subject in the normal sense, there is something that *would have been a subject* had it been uniquely embedded. For example, a narrative that had it been produced in a single environment would have been a narrative of entities in that environment, but because of the nature of its genesis (e.g. a historian who was unwittingly exposed to a superimposition of Earth and Twin Earth) it is not a narrative of anything.

The idea can be captured with the distinction between wide and narrow content. On this view, the multiply-embedded brain would be dealing with *narrow* contents, and hence there would be a single *narrow-content-structure*. And in the same way that narrow contents are sometimes understood to be functions from contexts to contents, so too this narrow-content-structure would be a function from contexts to (wide-) content-structures. But, the proponent of this position might argue, since there is not a single context, the function simply isn't correctly called. It is a one-place function into which two inputs are being crammed. The function doesn't get off the ground. Had the structure been uniquely embedded, there would have been a genuine narrative or a genuine content-structure – a wide-content-structure – there to define a subject. But as things stand, this didn't happen.

The contentualist might speak by extension of *narrow-subjects* and *wide-subjects*. A narrow-subject would be, on this view, defined by a narrow-content-structure. It is a function from contexts to content-structures, that is, a function from contexts to *subjects*. And a wide-subject (i.e., a subject as normally understood) would be what results when one has a uniquely embedded content-structure. And the contentualist who is a proponent of this view might claim that while in the Janus case there would be a narrow-subject, there would be no wide-subject – that is, no *actual* subject. Note that at least according to a contentualist, such narrow subjects would not be located anywhere, neither spatially nor temporally. The narrow-content-structures do not have any suitable demonstrative or indexical contents to define a location in space or time. If the self is defined completely or partially in terms of a narrative, model (self-model or

world-model, or both), or map, then there is trouble in such a case as this. The map/model/narrative constructed by my actual brain concerns me, my environment, my body. And Rico's concerns him, his environment, his body. And this is because the mechanisms that supply the suitable (wide-)contents are in place. But what sort of subject can be defined by a narrative that is about nobody, or by a map of nowhere, or by a model of nothing?

This is the second time the notion of an abstract subject has emerged. In the Scatterbrain case we met with *abstract processing time* and the possibility of a subject whose mental life was temporally abstract because it was not embedded in a unique actual temporal interval. The sort of abstract subject under discussion here is different and arguably more extreme, in that not only is the temporality unanchored, so too are all other contents.

2. There is one subject, but its world is a superimposition of two distinct locales. I suspect part of what makes this appealing is that it seems easy to imagine oneself into the position of this subject. We often experience things that overlap or are superimposed.

The brain-lubber will probably embrace this position. On this view, the function from contexts to contents is in receipt of a legitimate input, namely, a superimposition of the objects in the two environments. Notice for now that many, perhaps the great majority, of Janus's beliefs will arguably be false on this view. If asked, she will say and believe that she is holding one apple, but in fact she is holding two.[18]

3. There are two distinct subjects sharing a single brain, because there are two numerically distinct content-structures, owing to their different causal connections to different environments. On this approach, there is one narrow-content-structure. This is a function from contexts to contents, and this function is correctly called twice. Each of the two times it is called, it maps a context to a content-structure. (Note the difference between (a) one input consisting of a superimposed pair of contexts, and (b) two distinct inputs, each consisting of a single context.) On this approach, since there are two wide-content-structures we have two subjects – two *Jani*, so to speak. There are two qualitatively identical but numerically distinct narratives (one concerning events

[18] Maybe, it gets tricky. Since "one" on the lips (or brain) of such a subject might be analyzed as meaning *two*. I'll not pursue these nuances.

on Earth, the other concerning events on Twin Earth): two models, two maps, two sets of discursive commitments/entitlements, and so forth.

The point is that if what brains are doing is creating abstract subjects (temporally abstract, or abstract in the narrow-content sense) that are functions from contexts to genuine concrete subjects, then when there is *one* embedded context-supplying embodied point of view (even if there are two brains participating in its construction), there is *one* subject. But what happens when there are two? There are three possibilities: the process is an aberration and no contents are produced; or a content structure is produced, but it concerns a superposition of two environments; or two distinct content structures are produced in the same vehicle.

Note that while these options are here being raised in terms of the wild thought experiments I'm exploring, the question is one that could be asked of anyone who endorses a causal theory of content, since it is a real possibility that a given psychological or neural state might be caused by two numerically distinct but qualitatively identical objects. And since this is a manifest possibility even in mundane cases, the question is one that ought not be dismissed out of hand even by those who might find my thought experiments too outlandish to take seriously.

Now back to the main line of discussion. The one-subject-per-brain proponent has another line of defense for their counting policy. It is intuitions about what would happen in cases of causal divergence. The thought experiments described above incorporated the assumption that the two brains (jointly embedded in a single environment) and the two environments (in which a single brain is embedded) evolve in identical ways. But what if there is divergence? Let's take each case in turn.

What if the two brains embedded in one environment – the Raid brains – start to diverge? They run identically for some time, but then at some point one brain surrenders to the desire for a sixth slice of bacon and the other heroically suppresses the urge. If the motor commands are combined by an AND gate, then the body will not reach for the bacon, since only in cases where both brains send a motor command are they passed to the body. So from the point of view of the heroic brain everything is as expected – the arm will not reach for the bacon. The indulging brain, on the other hand, will be surprised and frustrated, since its motor command failed to move the arm. And at that point the differences will quickly snowball. The indulgent brain will panic or at least be puzzled, and start trying all sorts of movements, none of which will

work. And the heroic brain will then likewise be ineffectual, since the commands it issues won't, except by sheer luck, be matched by the other brain and hence implemented by the body. If the motor commands were combined with an OR gate, or by averaging, the bodily behavior produced by the divergent brains would be different but equally dysfunctional.

Because post divergence, it seems compelling to say that there are two subjects (call them Raid-A and Raid-B) this can seem like a pretty solid reason to claim that the two brains *before* the divergence were each supporting distinct subjects who were just happily in sync.

But this conclusion is not forced. A pure contentualist will claim that upon divergence, two subjects emerged where before there was one. That is, Raid-A and Raid-B are continuations of a single progenitor, *Raid*. Owen Flanagan (1992) has provided a sort of Dennettian contentualist account *very roughly* along these lines as an analysis of multiple personality disorder. The differing views here (one subject splitting into two vs. two all along) are analogous to the views under discussion in David Lewis's proposal concerning Parfittian fission (Lewis 1976). Lewis, recall, resisted the idea that in fission cases we have one person becoming two, and instead proposed that in cases where fission occurs, there were always two persons there to begin with. In the present context, the contentualist is siding with the one-subject-fissioning-into-two camp, and the one-subject-per-brain proponent is siding with Lewis.

What about the Janus? Suppose that the brain indulges in the sixth bacon strip, but as it reaches out to pick it (them) up it slips out of the hand of one of the bodies, but does not slip out of the other. Clearly surprise and confusion will follow. Depending on how the sensory signals get combined, the brain will be presented with either very little (if combined by an AND gate), or too much (if an OR gate), or a structure-dissolving fusion of both environments (if averaged). To keep it easy, let's assume that they are combined with an OR gate, so that what gets presented to the brain is something like a superimposition of both bodies' sensory deliverances. In this situation it is easy to imagine that one is *the* subject, a subject that is now presented with an *obvious* superimposition of two different environments. I say "obvious superimposition" because, on this line of thought, this single subject has always been presented with a superimposition of the two environments, but this fact was not obvious, since the two environments were identical. On this view, there was only one Janus before the

divergence and only one after. But after, Janus is made aware of the superimposed nature of her experience.

But while this is one natural way to conceive of the situation, it is not the only way. The pure contentualist must maintain that there were two Jani – Janus-A and Janus-B – before the environmental divergence, and that both of them remain after the divergence. So if I am Janus-A, then on this line of thought, pre divergence I was a subject normally embedded in my environment – Earth, say. And my mind was in the business of contemplating contents anchored to Earthly entities. And the brain supporting my thoughts was also supporting the thoughts of Janus-B, who was dealing with Twin Earth.

But post divergence the brain supporting me became subject to causal influences (from the body on Twin Earth) that disrupted its ability to support me as a subject in the same way. Of course my brain was always subject to causal influences from Twin Earth, but until the divergence, these causal influences did not disrupt the brain's ability to support me as a subject. But now the causal influences are disruptive. My mind is still in the business of dealing with contents anchored to Earthly entities, but business is not good. I see a hand grasping the bacon, but I also see a superimposed image of an empty hand over a falling bacon slice. The subject on Twin Earth would be analogously disrupted, beset by hallucinations caused by Earth's influence on his brain. This is a very tricky and slippery position, and I'm not sure it will ultimately hold up. But rather than get bogged down here, I'll just consider the last couple of paragraphs as written in pencil, register the fact that the pure contentualist has a significant (though perhaps not insurmountable) challenge here, and move on.

Let's take stock, since this section was fairly viscous. Unlike the manipulations in Sections 1, 2, and 3, which hinged on assessing how the spatial and temporal features of the mind would, according to various theories, be affected by manipulating the spatial and temporal features of its physical implementation, we are now concerned with assessing the contribution that the physical implementation makes in the *individuation* of minds. The manipulations of this section involved exploring the differing contributions made by brains and embedded bodies to minds' synchronic individuation. To my knowledge this has not been explicitly explored in the literature. Among other things, it raises an issue for functionalists concerning the individuation of *functional systems*. Functionalists have been concerned with individuating functional states within

a system – pain states and belief states, for example, have different functional profiles within a functional system. But on this view, a *mind* will be precisely a system of functional states, and as such, individuation questions can be raised concerning such systems as a whole. I've sketched two ways functionalists might respond, dirty and pure functionalism.

One initial response will be to appeal to the physical states implementing the functional system: *dirty functionalism*. This approach faces challenges, including (a) the fact that it constitutes a significant concession to precisely the approach it was designed to defeat (physicalism/identity theory), and (b) the fact that it seems definitionally incapable of accommodating the possibility of hardware-redundant functional systems – and such systems seem to be manifestly possible. The good news is that this approach, because it makes the concession concerning the importance of the physical implementation, is able to say things about the scenarios involved in this section that are intuitively appealing – namely, that there is always one subject per brain.

The second response is pure functionalism, which lets the input/output interfaces do the individuative work for functional systems. This is definitely more resonant with the original motivations of functionalism than letting the material implementation do this work. And the interface is the embodied/embedded point of view, so on this approach, the number of subjects will track the number of embedded/embodied points of view. The hardware implementing the functional systems is important only insofar as it is up to the task of maintaining the right functional system between the inputs and outputs. And success at this job is entirely consistent with two brains jointly supporting a single subject (Raid), or one brain supporting two (Janus).

The lessons for the contentualist are similar, though they are forced by slightly different considerations. The dirty contentualist will pattern like the dirty functionalist and allow the physical implementation of a content structure to do individuative work. But in the case of the *pure* contentualist, the embedding environments do the individuative work because these are what anchor the demonstrative and indexical (and other) elements of the content structures that define the subject. That is, the pure functionalist appeals to the embodied/embedded point of view (the interfaces, the body), whereas the pure contentualist appeals to the embedding environment that is accessed via those interfaces. But either way, the brain is deemphasized and the body/environment are given prominence.

I want to sum up the functionalist and contentualist lessons a bit more contentiously. I think both groups have been skating for the last few decades on ill-gotten plausibility. They both distance themselves from lubberish views – and rightly so. But this distancing has taken the form of highlighting manipulations that swap the brain basis out for some other sort of basis (Martian brains, for example). But one thing all of these basis-swapping manipulations have in common is the assumption, shared by the lubbers, of a one-to-one mapping between subjects and implementing entities (brains, computers, or whatever). The lubbers can legitimately presume upon this assumption because the importance of the brain itself is built into their view. But the functionalists and contentualists cannot presume upon this assumption. There is nothing about functionalism or contentualism that entails a one-to-one mapping between independently individuated bases and minds. Basically, by not facing these possibilities they have been drafting the lubbers on this – and not legitimately so. As soon as these questions are raised, the fact that subclinical lubberish leanings have been in place all along becomes visible in the form of the drive towards dirty functionalism and dirty contentualism.

My point is not that functionalism or contentualism (my own view) are mistaken because they have these counter-intuitive consequences. Rather, I think the functionalist and contentualist should remain pure, and face the fact that their respective views have some consequences they hadn't recognized. We didn't really think that exorcising lubberism was going to be that easy, did we?

Another issue that got raised is the possibility of something characterizable as an abstract subject. In Section 3, the abstract subject was the lubber's gambit to save a mental timeline for Scatterbrain. In this section the possibility of an abstract subject emerged for the contentualist as a way to describe a situation in which something appears outwardly to be a normal functioning subject – as both Janus-A and Janus-B would appear to people on Earth and Twin Earth – despite the fact that their shared brain might be manipulating only so-called narrow contents. I argued in Section 3 that the temporally abstract subject Scatterbrain would need to be embedded in a specific temporal interval defined by the embedded/embodied point of view to avoid being characterized as having only an abstract mental timeline. In this section, the issue is one of the information-processing system's being embedded in a unique environment in a way capable of bestowing genuine broad contents on the relevant, mostly locative,

elements of the content structures. I will return to the topic of abstract subjects in the final section.

The upshot has been that the brain-lubbers have probably the most initially appealing position: one subject per brain. The BBE-lubbers have a counting problem. The functionalists and contentualists are faced with a crisis of faith: either remain pure and accept the counter-intuitive consequences to the effect that a single brain can host two minds and vice-versa, or renounce purity and allow physical stuff to play a crucial individuating role, a role that it typically does not have on functionalist or contentualist views and that seems more consonant with lubberish approaches.

## 5. Diachronic Identity: *Bourdin*

Suppose, as in the Janus and Raid scenarios, that a subject is the result of cloning, but this time we keep both brains and both bodies. Each body is in its own environment, one on Earth and the other on Twin Earth. In such a case, the brain-lubber, the functionalist, and the contentualist can agree that there are two normal subjects with qualitatively identical (narrow-content) experiences. We have $\text{brain}_1$ interfaced with $\text{body}_1$ in $\text{environment}_1$, and $\text{brain}_2$ interfaced with $\text{body}_2$ in $\text{environment}_2$. Since the two brains process information in the same way, and since the bodies and environments are qualitatively and dynamically identical, their thoughts, feelings, and actions, will be numerically distinct but qualitatively (narrowly) identical.

But let's suppose the interfaces between the brains and bodies switch back and forth. Consider an interval between $t_1$ and $t_2$, and call it $\text{interval}_1$; and likewise call $\text{interval}_2$ the interval between $t_2$ and $t_3$; and so forth. During $\text{interval}_1$ $\text{brain}_1$ is interfacing with $\text{body}_1$ and $\text{brain}_2$ is interfacing with $\text{body}_2$. But at $t_2$ there is a switch in interfaces such that during $\text{interval}_2$ $\text{brain}_1$ will be interfacing with $\text{body}_2$ and $\text{brain}_2$ will be interfacing with $\text{body}_1$. This switch happens very quickly – so quickly that it makes no impact on the functioning of the underlying causal operations. For $\text{interval}_3$, the interfaces switch back to the same configuration that had during $\text{interval}_1$. And let's suppose that these switches happen with some frequency, perhaps once every second. We can call this subject *Bourdin*, after the infamous serial impostor Frédéric Bourdin.

We have bodies embedded in environments, but the brains that are causally running the bodies are being swapped back and forth. As I am conversing with what I take to be a normal subject about skepticism, for instance, the brain that caused the body in front of me to utter "I think," is different from the brain that caused the phonological continuation "therefore I am."

Keep in mind that the question now is not one of the *location* of the subjects. In Airhead we explored manipulating the location of the brain with respect to the location of the embedded point of view, and the discussion concerned which of the two the location of the subject would track. Now we're looking into an issue that is, in some respects, deeper. Deeper not so much for the lubber, who will still identify the subject in the current scenario with the brain, both in terms of location and also, in the current scenario, diachronic identity. But for the functionalist and contentualist, even if you think a subject's location is determined by the location of the interfaces, the question remains, *which subject is it* whose location is being determined by the interfaces? Is the subject's diachronic identity tied to that of the brain (even if, as the contentualist and functionalist would have it, *where* that subject is located at any time will be determined by the location of the embedded point of view that that brain is causally connected to)? Or is it the case that the embedded point of view determines not only the location of the subject at any time, but also provides for the diachronic identity of the subject over time (regardless of whether the computing machinery supporting that point of view is getting swapped out)? During my conversation, am I speaking with the same subject in $\text{interval}_2$ who I was speaking with during $\text{interval}_1$? All can agree that I am interacting with the same body in the same environment during the two intervals. But the brain that is behind the causal curtain is different. Which of these determines the identity of the subject?

The brain-lubber will of course maintain that there are two Bourdins, one per brain, and that these Bourdins are being alternately interfaced with different bodies in different environments. I am conversing with $\text{Bourdin}_1$ during $\text{interval}_1$, and with a different Bourdin, $\text{Bourdin}_2$, during $\text{interval}_2$. The brain lubber *might*, if she is also a content externalist, claim that the subjects' minds in this case are … how to put it … *messed up*, perhaps even unable to grasp coherent thoughts. Here is the line of reasoning. Suppose the lubber's views on diachronic identity are right, and suppose also that I am the subject implemented in $\text{brain}_1$, and that I am peeling an apple. I ($\text{brain}_1$) will, during $\text{interval}_1$, be processing visual information from $\text{apple}_1$ in the hands of

$body_1$ on Earth. But during $interval_2$, I will be getting qualitatively identical visual input from $apple_2$ held by $body_2$ on Twin Earth. While it will seem to me that I am just peeling a normal persisting apple, in fact, the thought "that apple" has no clear referent. There is no single object that I have been tracking over time. And on at least some externalist views of perceptual content, this would render the thought-attempt an aberration. Depending on how often the switches occur, it might be questioned whether any coherent contents are grasped at all. If the switch happens only once a day at 3 am while I am sleeping, then I would have a full day of consistent interaction with a single environment. I would be mistaken that *this* computer is the same one I was typing on yesterday, but I would now still be able to have a thought about "this computer", it seems. But if the switches happen once every second, or even more frequently, then it starts to be difficult to make sense of my grasping of any coherent (broad/externalist) perceptual contents at all. If there *were* an object of my thought, perhaps it would be *at best* a spatially discontinuous 4-dimensional apple-worm (so to speak) composed of alternating time-slices of the two conventionally individuated apples.[19]

The BBE-lubber is in a tougher position. In addition to the same problem about grasped contents there is a hardware issue. During $interval_1$ everything is normal: we have $brain_1$ interfacing with $body_1$, which in turn is causally embedded in $environment_1$. And recall that for BBE-lubbers the mind pervades at least two, perhaps all three, of these. But what happens during $interval_2$, when we move from $brain_1$ being embedded in $body_1$ to being embedded in $body_2$? The sticky bit is that all the components required for mindedness are present – a brain embedded in a body, which is embedded in an environment. Does the pre-switch mind supported by *the combination* of $brain_1$ embedded in $body_1$ go out of existence to be replaced by a new mind, composed of $brain_1$ embedded in $body_2$? Does the old mind pop back into existence when they switch back? Or perhaps the original mind is still operational, but now just dispersed over causally isolated hardware? These question will be explored in the next section.

I'm really not sure what a functionalist would say, though I can see a few ways of going. A pure functionalist will probably stick with an individuation of functions that doesn't appeal to physical implementation, but rather appeals to the identity of the input/output interfaces. In particular, she won't worry about whether the physical stuff

[19] Notice how much more radically mistaken I would be in this case than in other cases of external content determination, for example elms and arthritis.

that is implementing the function gets swapped back and forth. Each mind (= system of functional states) stays put in a single environment despite the switching of the neural hardware behind the curtain. After all, that's textbook functionalism. You can change the stuff that implements the functional states without changing the functional states themselves. And the mind is a system of these functional states.

In the last section we saw the dirty functionalist insist that the physical entities implementing functional states were at least partially individuative of those functional states, and by extension of the system of functional states that constitutes a mind. The issue, recall, concerned *synchronic* individuation – whether two brains can implement one mind (or vice versa) *at a time*. It would now be open to a dirty functionalist to claim that physical implementation is a factor in diachronic identity as well. On this view the functional systems, and hence the subjects they implement, are being swapped back and forth with the brains.

Two more positions in logical space are (a) to claim that physical implementation does not matter for synchronic individuation (this, recall, was the *pure functionalist* position) but that it does matter for diachronic identity. On this view, while one brain might implement two minds (synchronic individuation *not* determined by physical implementation), those two minds would have their diachronic identity tied to that brain. And (b) to claim that physical implementation does matter for synchronic individuation but not for diachronic identity. It is not obvious to me what the motivation would be for the functionalist to adopt either of these locations in logical space, though. From the standpoint of functionalist commitments, it would seem to make sense to take synchronic individuation and diachronic identity to pattern similarly with respect to their dependence on physical implementation. If you're going to be dirty, just keep it dirty; and if you're going to be clean, stay on the straight and narrow.

On all the above approaches, there will be an issue about the contents figuring in these functional systems, which I will discuss below in the context of the dirty contentualist. And of course dirty functionalists who think the body and/or environment are also crucial parts of the functional system will also face problems analogous to those of the BBE-lubber discussed above (and discussed in much greater detail in the next section).

The dirty contentualist, like the dirty functionalist, will be inclined to allow diachronic identity to be determined by physical implementation, just as with synchronic individuation. But the pure contentualist's situation is trickier than that of the pure functionalist. There is no obvious compelling reason for the pure functionalist to place different weight on the contribution of physical implementation to diachronic identity than was placed on its contribution to synchronic individuation. The material basis of functional states could swap in and out over time just as easily as they could be redundant or do double duty.

But the pure contentualist is in thicker tar. One issue is that there are two kinds of pure contentualist which were not necessary to distinguish until now. All contentualists take mindedness to hinge on contents, but there are two ways to understand how contents come about. Some, like Brandom, take contents to be a matter of attribution by third (or second) parties, while others take contents to be generated in some manner that does not require any other agents or subjects attributing intentionality or anything of the sort. The key difference is, to put it in blunt terms, whether the stuff that is responsible for generating the content is "internal" to the subject in the sense of being independent of anyone else's stances or attributions.

Let's consider attributional contentualists first. According to such a theorist (Dennett and Brandom would be examples), the contents grasped or entertained by a subject are determined by other people adopting a stance towards that subject. So for instance because I am the one who is interpreting my interlocutor and attributing contents, when the embodied agent in front of me says "I think" and then follows it up with "Therefore, I am", I am attributing content attitudes to a single persisting agent.[20] The attributors aren't swapping back and forth. They are part of the persisting environment. And the thing they are attributing the contents to is the embodied agent in front of them. It isn't clear whether on this view there is any room for any position other than that the behind-the-scenes swapping of the supporting hardware is irrelevant. The subject's diachronic identity is determined by the diachronic identity of the content-structures involved, and these are being attributed by content attributors

[20] To make the case for this in Brandomian terms: if during $\text{interval}_1$ the bodily agent in front of me says "If it is raining, the sidewalks will be wet" and then during $\text{interval}_2$ that bodily agent says "It is raining," then I will treat that very bodily agent as committed to the claim that the sidewalks are wet.

in the environment to an embodied agent they are interacting with in that environment.

Pure contentualists who take it that the contents manifest non-attributively could hold that *vehicles carry their contents with them*. During $\text{interval}_1$, $\text{brain}_1$ has physical states that are the vehicles of representations in virtue of being causally impacted by objects in $\text{environment}_1$. And maybe this is supplemented with some capacity that the neural systems themselves have for constructing or creating contents in some internalist way. If vehicles carry their contents with them, then during $\text{interval}_2$, when $\text{brain}_1$ is interfaced with $\text{body}_2$, that brain carries with it those (broad) contents that were initiated in $\text{environment}_1$ during $\text{interval}_1$. And since the subject is a structure of such contents, the subject follows the brain. On this view, though I would take it to be the same interlocutor who is making successive statements, in fact, the interlocutor that made the first statement is not the one who made the second.

## 6. Existence and Recombination: *Theseus*

Now things are going to get a little weird.[21] Let's start with the supposition that I am the result of a cloning procedure that has resulted in two qualitatively identical brains and two qualitatively identical bodies that have been placed in qualitatively identical environments on Earth and Twin Earth, respectively. Each of the brains is divided into five parts (say) with the causal interfaces to and from all the other parts of that brain maintained via transceiver linkages. This is similar to Scatterbrain, though now we have two brain+body subjects involved, and we are not concerned about temporal displacement. So far this adds nothing new that wasn't present in prior sections.

Let's label each of the parts of brain $b_1$ as $b_1[a]$, $b_1[b]$, $b_1[c]$, $b_1[d]$, and $b_1[e]$. Similarly, the parts of brain $b_2$ we can label as $b_2[a]$, $b_2[b]$, $b_2[c]$, $b_2[d]$, and $b_2[e]$. These parts might be something along the lines of the left hemisphere, right hemisphere, brain stem, cerebellum, and midbrain. Accordingly, a complete causally-integrated brain

[21] That was a joke.

(CCIB for short) will consist of five components, one each of parts [a]-[e], with their causal interfaces hooked up in the correct way.

The brain parts can now be mixed-and-matched. If we have relays set up at the interfaces, we can swap parts: for example, we can unplug $b_1$[a] (the cerebellum, say) from the rest of brain $b_1$, and swap in $b_2$[a] (the cerebellum) from brain $b_2$. Since the swapped-in component is functionally/causally identical to the one swapped-out and has plug-compatible interfaces, and since we are making these swaps on a time scale far below that of the relevant neural causal mechanisms, brain $b_1$ will remain a CCIB. (Though of course whether we should still call this brain '$b_1$' is an open question since only 80% of it is the same neural tissue as before the replacement. But that vanilla identity issue is an irrelevancy for present purposes. This topic will emerge later in this section.) And we can fill the void left in brain $b_2$ – the void created when part $b_2$[a] was plugged into the vacancy in brain $b_1$ – with the corresponding part from brain $b_1$. Imagine that the various parts of the brains are mixed-and-matched at intervals, with the stipulation that throughout any interval there will be two CCIBs consisting of one each of parts [a]-[e], with one CCIB interfaced with each body. Because we are assuming that each brain part will function identically if given the same causal inputs and also that each environment will provide the (qualitatively) same inputs, each brain part will always be in the same biological/causal/functional state as the other corresponding part. I will describe this by saying that the brain parts in question are *in sync*. This means that when synced parts are swapped around, everything proceeds causally and functionally just as it would have proceeded had the switch not occurred.

Recall MPH, the mind-preservation hypothesis. This was the hypothesis that removing a brain, or even spatially or temporally dispersing it, would not imperil the mind or subject supported by that brain, so long as all the relevant causal (and maybe chemical) interactions were maintained through suitable interfaces. Whether MPH applies to cases of CCIBs with parts being switched around is going to be an open question in this section. One who accepted MPH in all the previous manipulations might well feel a pull to reject it now, since this is the first manipulation in which the parts of a single brain aren't filling their roles by remaining in causal contact with all the other parts of that same brain.

But recall MPH's weaker sibling, the *behavior-preservation hypothesis* (BPH): so long as a body is being operated by a CCIB (in which all swapping parts remain in sync), the overt *behavior* manifested by the embedded body that this CCIB is

interfacing with will be preserved, meaning it will be the same as it would have been had the swapping not occurred – indeed, as if no brain interfacing procedure at all had been implemented. The idea is that since all the same causal structures are in place, a swap won't have any overt consequences. In order to deny BPH, one would have to accept the idea that, at the macro level, identical causal structures can produce different causal consequences, since all these causal structures are, *ex hypothesi*, unchanged. How one might deny BPH in this situation without courting the paranormal is unclear to me.

With that piece on the board, let's start with the easy things to say. The pure functionalist and pure attributive contentualist will count two subjects, call them $\text{Theseus}_1$ and $\text{Theseus}_2$, each persisting in one of the environmentally embedded bodies. They will both agree that MPH and BPH are applicable. The reconfiguring infrastructure is irrelevant. The reasons for this should be familiar by now.

The situation for everyone else – dirty functionalists, dirty contentualists, pure but non-attributive contentualists (purity is relative only to the synchronic individuation issues of Section 4) and BBE-lubbers – is far more complicated. Let's not worry about contents just now. Let's focus on brains, since for all of these theorists, subject diachronic identity is tied to physical implementation. They have different reasons for thinking so: BBE-lubbers just identify the mind and brain(/body/environment); dirty functionalists take diachronic identity of functional states to be dependent on diachronic identity of physical implementation; and the relevant contentualists maintain that contents are carried, diachronically, by their vehicles.

Suppose that things start out at $t_1$ with $\text{Theseus}_1$ interfacing with his original biological brain which, while spatially dispersed, is still a CCIB. That brain, body, and environment have all been together for his entire life up to $t_1$. Things are relatively normal. At this point $\text{Theseus}_1$ is Scatterbrain. So MPH and BPH are arguably both in play. Then at $t_2$, two or three of those parts are swapped out to the other brain, and its corresponding parts swapped in. What happens to $\textit{Theseus}_1$ on the views currently under consideration, those for which mentality follows, for one or another reason, the physical implementation? There would seem to be at least the following options:

1. $\text{Theseus}_1$ persists through $t_2$, but the five parts that compose his implementation base at $t_2$, the five parts of brain $b_1$, are no longer interacting together as part of the same causal whole. They are divided between two separate causal systems now. The

idea here is that minds persist and are always tied to the same five (in this case) "original" biological brain parts, regardless of whether those parts continue to causally interact with each other, so long as each is still functioning normally embedded in its own synced CCIBs. On this view, there will always be two minds in this scenario, the same two that started out. But they will have their implementation bases (original biological brain) distributed across different CCIBs and bodies at different times.

This might be the right thing to say, but it seems that it should seriously lack plausibility to anyone who holds one of the views currently under consideration (views that take physical implementation seriously in mind identity and individuation). There is considerable intuitive pull towards the idea that the continued persistence of any particular mind, on these views, should depend on the continued causal interaction of the *original* brain parts *with each other*.

2. Theseus$_1$ goes out of existence. The idea here is that Theseus$_1$ is (as in (1) above) tied to a certain collection of brain parts, the original brain say. But on this option if these brain parts stop interacting *with each other*, then even if they are still individually functioning in synced CCIBs as they would have functioned had they remained connected together, they stop supporting *Theseus$_1$'s mind*, and this mind pops out of existence.

The position that this mind goes out of existence when its original brain is no longer its own CCIB leaves two unanswered questions. One is what happens to this mind if those original brain parts reconnect? Does it come back into existence? Or, once eliminated, does it stay eliminated even if the original brain reunites? The second question is if this mind goes out of existence (*either* because the original brain is currently shuffled into distinct CCIBs, *or* because it *was* so shuffled, but reintegrating it didn't bring me back), then what, if anything, *is* there? No mind at all? A new numerically distinct mind? Let's tackle the first question first. There are two possible answers:

2a. Once Theseus$_1$'s brain parts have been disconnected and distributed across CCIBs, then even if they are rejoined into their original configuration at $t_3$, this mind is still gone. The initial disconnecting of Theseus$_1$'s brain parts annihilated him, and he can't come back. There are reasons to resist this. Recall the Wada procedure, in which one of a person's hemispheres is anesthetized while the other remains awake.

Regardless of whether you think the person persists when one of the hemispheres is asleep, everyone (as far as I can tell) agrees that when the sleeping hemisphere reawakens and resumes its normal interactions with the other hemisphere, the person is either *back*, or is *still there* because they never left. This seems like the right thing to say, and it is quite similar to the reconstituted CCIB case envisioned here. This suggests:

2b. When the original brain parts are reunited at $t_3$ into their original CCIB, $\text{Theseus}_1$ comes back. Notice something interesting about this proposal. Consider the temporal interval from $t_2$ to $t_3$, during which CCIBs other than the original brains ($\text{Theseus}_1$'s and $\text{Theseus}_2$'s) were embedded in the two bodies. Both bodies were interacting *normally* with their environments, as per BPH. Suppose that during that time after the initial swap, when $\text{Theseus}_1$ is, *ex hypothesi*, out of existence, a phone call came in and was answered by his body. When the original brain gets reunited and he pops back into existence, he will presumably have the memory (actually, a quasi-memory, see Shoemaker 1970) of having responded to a phone call.

Why think he would have this (quasi-)memory? Two reasons. First, by hypothesis, all the physical states that would normally support short- and long-term memory are in play. The brain part with *his* hippocampus was at the time of the phone call part of a CCIB that received a phone call and operated causally so as to produce the overt behavior of responding to the call. This action presumably left physical traces of precisely the same sort (but for the different causal source, if at that time that part was interfacing with Twin Earth, say) that would have been left had the parts never been swapped out. So on this position, even though $\text{Theseus}_1$ just popped out of existence at $t_2$ and popped back in at $t_3$, *it would seem to him that he had been normally persisting through all intervals*.

The second reason for thinking that $\text{Theseus}_1$ would have a memory of having responded to the phone call hinges on BPH. If after he pops back into existence someone asks him "Did you get a phone call?" his body will, by BPH, produce the response "Yes." If he has no recollection of the phone call, then he will immediately be shocked to hear his body produce a sentence he doesn't believe or feel like he intended to say, and then a lot of unusual behavior that violates BPH will ensue, such as him uttering the sentence "Why did I just answer 'Yes'?" which is not the behavior that would have emerged had the swapping never occurred.

Anyway, the point for now is that on this view, if $\text{Theseus}_1$ comes back into existence, he will have no recollection of having been wiped from existence and will rather feel as though he has persisted normally throughout the entire interval.

Now, let's discuss the possibilities, on the view that his mind goes out of existence when his original brain has swapped parts, concerning what *is* in existence during those intervals. I will list these options as (2i), (2ii), and (2iii), since this issue is orthogonal to the options listed as (2a) and (2b):

2i. When any configuration of a CCIB that does not consist of exactly the parts that composed one of the original undetached brains is present, that CCIB is supporting no mind or subject at all. During those intervals, the bodies are functioning normally from an overt behavior standpoint – BPH is satisfied – but there is no subject or mind present. MPH is not satisfied. Note that anyone who thinks that philosophical zombies are not possible will have to deny that 2i is an option and will in general have to maintain that MPH holds whenever BPH holds, since to deny that *just is* to accept the possibility of philosophical zombies. So if you think that philosophical zombies are not possible, then you will need to accept the idea that any non-original CCIBs support genuine minds.

2ii. When any configuration of CCIBs that does not consist of exactly the parts that composed one of the original undetached brains is present, that CCIB is supporting its own proprietary mind/subject – one supported by, and only by, exactly that specific configuration of brain parts. This has consequences as well. With two brains divided into five brain parts, there are $2^{(5\text{-}1)} = 16$ different configurations possible. Are we to take it that there are 16 distinct minds, popping into and out of existence as the parts are reconfigured? If they are equiprobable, then $\text{Theseus}_1$ is actually around for approximately 6% of the time, though when ever he is around he is under the impression that he is a continuously persisting self, unaware of the fact that 94% of his remembered past is mere quasi-memory, and concerns events that occurred when he did not even exist.[22] We would have two brains worth of neural tissue supporting (one at a time) 16 distinct minds, all of which believe themselves to be persisting continuously.

[22] This is the best case. For anyone who chose (2a), presumably every subsequent configuration, even if it is a configuration that was already implemented, houses a new mind, in which case 100% of each short-lived mind's recollection would be quasi-memory.

2iii. Mind persistence is a matter of degree, depending on the configuration of brain parts. The idea would be that if, between $t_2$ and $t_3$, the brain operating the body on Earth is 60% Theseus$_1$'s original brain and 40% Theseus$_2$'s original brain, then the subject present on Earth during that interval is a 60/40 mixture of Theseus$_1$ and Theseus$_2$.

This is essentially a fuzzy version of 2ii, a proposal consistent with Parfittian ideas concerning personal persistence through time (Parfit 1971). It is just letting identity conditions for the supported minds hinge on identity conditions for the supporting brains. There is on this option, as with 2ii, something uniquely associated with each CCIB configuration – a proprietary fuzzy mixture of the original minds, as opposed to completely novel minds. There will still be 16 possibilities, though. The 60/40 mixture one gets by swapping out parts [a] and [b] will not be the same as the 60/40 mixture resulting from swapping out parts [c] and [d].

Let's take stock. So long as the causal structures are in place to keep the embedded point of view functioning, pure functionalists and pure attributive/interpretive contentualists will ignore the bedlam of the neural shell-game and reckon one continuing subject in and per embedded point of view.

Lubbers of all stripes, dirty functionalists, and all non-interpretive/non-ascriptive contentualists have a number of options. But they involve accepting at least one of the following:

A. The idea that a mind can continue to exist even when the supporting brain parts are no longer interacting with each other, but are rather interacting in different CCIBs (option (1) above.)

B. The possibility of philosophical zombies. (Options (2a) and (2i) above.)

C. The possibility of a large number of distinct minds, all supported by two brains worth of neural tissue. All of these minds pop into and out of existence, perhaps very quickly, each under the mistaken impression that it is persisting through time, when in fact most of the time they don't exist and their body is inhabited by other minds in a series or in rotation. (Options (2ii) and (2iii) above.)

## 7. Discussion

The thought experiments that have driven the discussion have been extreme, and for some readers who don't want to follow the various threads I've explored, this might seem like a convenient excuse to dismiss them. But I don't think their extremity is a valid excuse for dismissal. For several reasons. The explorations here are explorations concerning what minds are. And on a common measure of fundamentality, if we have two theories of phenomenon X, and $\text{theory}_1$ holds in a wider range of circumstances than $\text{theory}_2$, then $\text{theory}_1$ is more fundamental than $\text{theory}_2$. Theories of mechanics and dynamics that hold only at modest relative velocities are less fundamental than those that hold both at modest velocities as well as at velocities approaching the speed of light. A defense of Newtonian mechanics against relativistic mechanics on the grounds that the defender found the possibility of accelerating anything to near-light speeds outlandish would, I hope, be seen to be lacking. Similarly, theories of particle physics that hold only at low energies are less fundamental than those that hold at low and very high energies. Similarly, it strikes me as a good principle that a theory of what minds are that holds in a wide range of circumstances – even if those circumstances seem outlandish (today) – is superior to one that holds only in a restricted subset of those circumstances, and just packs up and goes home in less restricted sets of circumstances.

I should be clear that though the examples I've been exploring involve, so to speak, very high energies and velocities, the goal is to achieve a deeper understanding of what is happening in the everyday case. Our intuitions about physics and other domains reflect the restricted situations in which those intuitions formed, both at an individual and cultural level. Accordingly, exploring situations that help us reveal these limitations is valuable. I can't help but see reactions against exploring crazy thought experiments as precisely the wrong reaction. Such reactions are a fantastic way to guarantee that any intuitive blinders we may have stay in place.

A second reason that extremity shouldn't dissuade us is that it seems to me that the sorts of extremity present in the manipulations I've discussed is, in the relevant sense, metaphysically irrelevant. What I mean is this. Suppose I am about to take a few days off to actually flip a fair coin 10,000 times. We can all agree that it would be

unbelievably unlikely that I will get tails all 10,000 times. But does anyone actually think that anything metaphysically interesting hangs on whether this actually happens? Is anyone prepared to claim that if the coin comes up tails 10,000 times in a row, then mind/brain identity is true, but if it doesn't, then functionalism is true? Or that the metaphysical nature of minds could hinge on whether there are gravitational singularities? The invocation of extremely improbable or nomologically extreme situations often serve a useful rhetorical function – they can create imagined contexts in which the differences between two theories can be made manifest. But the idea that somehow appeal to unlikely or extreme thought experiments counts as "cheating" or sneaking something into the situation that surreptitiously changes the metaphysics of the situations seems wrongheaded to me – it's a too-convenient way to dismiss spotlights that are shining on places one would rather keep dark. At the very least such a claim would stand in need of defense in the particular case.

The only exceptions would be nomological or logical *impossibility*, and even there I'm inclined to be liberal and insist that the skeptic ought to be prepared to give some idea of why the lessons drawn from a description of a situation designed to make the differences between two theories manifest should be ignored, even if that situation hinges on some elements that are nomologically impossible. I should point out, though, that I don't think any of the thought experiments above involved any logically impossible scenarios, and with the possible exception sending information back in time, they also don't involve anything nomologically impossible. And so nomological impossibility would be a threat only to Rip and one of the Scatterbrain scenarios. Someone who wishes to dismiss these explorations would have to give some indication why the fact that the scenarios envisioned would be unlikely to occur or technologically difficult to implement should matter. And in fact, if my remarks about fundamentality are correct, then such situations are precisely the ones we should be seeking out.

That's all I have to say in defense of extreme thought experiments.

I think a number of interesting things have emerged from the above explorations. But I want to take a minute to highlight one in particular. It is what I believe to be a previously unappreciated affinity between two general approaches in the literature that superficially look very different: contentualism and embedded/embodied/extended approaches. As I pointed out in footnote 10, the former tends to look very hyper-intellectual, hinging on narratives and deontic scorekeeping structures and the like,

while the latter approaches tend toward the hypo-intellectual, and the hyper-visceral. But one thing that the extreme scenarios have made visible in the cloud chamber is the extent to which a workable contentualist program must be built on embodied/embedded foundations. In order for the content-structures to support anything recognizable as a mind, they must be built upon a bedrock of indexical and demonstrative elements that depend entirely on the embodied/embedded point of view. And the dependence ends up being quite strong, ultimately being the anchor points not only for the spatial and temporal locative contents, but the criteria of individuation and identity as well.

I imagine that some readers might well wonder whether what I am saying here is consistent with things I've said in the past. I mean this. I've hinted at several places in this paper that I consider myself a contentualist – probably a pure contentualist. And the pure contentualist is one who downplays the importance of the brain in an account of what the mind is, including its individuation and identity. The contentualist must see the embodied/embedded point of view as the material from which mindedness is built. The brain provides information-processing infrastructure that does the building. As that which implements the machinery of mindedness it is without doubt crucial. But its contribution must be understood in the right way.

How does this compare with stances I've taken in the past, for instance, in my paper titled "In defense of some Cartesian assumptions concerning the brain and its operation" (Grush 2003)? I take the views to be entirely consistent. My point in that earlier paper concerned the location of, so to speak, the machinery of mindedness. My opponent was someone who thought that the information-processing infrastructure of the mind had to be extended beyond the skull – and in the spirit of a sort of functionalism that claims the mind is a functional system implemented in the brain and environment, this opponent typically claims that it is the *mind* that is thereby extended. There are various reasons that folks have held that position, but none of them seemed compelling to me. I argued there that the normal human brain, internal to the skull, had the wherewithal to do what was needed to support mindedness. But I see this as an entirely distinct point from the point of this paper, which is that the thing that is induced, the mind that is created by the brain, is a thing that is defined by content-structures that have the embodied/embedded point of view as their content-providing bedrock – though not their information-processing bedrock. Indeed, I suspect that confusion between the two sorts of dependence is behind some of the positions in

the literature that I disagree with. But diagnosing all of that is beyond the scope of this paper. Let me just close this part of the discussion by saying that questions about the nature and location of the machinery implementing mentality should be kept as distinct from questions about the nature and location of the mind as we should keep questions about the nature and location of Hamlet from questions concerning the nature and location of copies of the play *Hamlet*.

I've said that I currently consider myself a contentualist, probably a pure contentualist. But this paper is not an argument for contentualism. I think Section 1 makes some solid headway in terms of defining and clarifying the contentualist position. But in terms of what happens in most of the paper, contentualism is one position among several that are tested in various thought experiments, and in some cases very counter-intuitive consequences of the view are revealed.

Given this, one might wonder what the conclusion of all this is supposed to be. This paper does not have a single overarching argument aimed at a specific conclusion. There are many arguments and many conclusions. Or if you prefer, the overall conclusion is that when a number of novel manipulations are considered, most philosophical positions on the relation between the mind and brain can be shown to lead to one or another implausible consequence.

Such aporetic musings will certainly leave some readers dissatisfied. This paper certainly does not have a standard structure, with a clearly defined thesis and an argument constructed to support that thesis. And these days, this is precisely what many philosophers identify (to our detriment, I believe) as the Platonic form of philosophical research output. So if there is value here, that is not where it lies.

I am reminded of a passage from Bernard Williams' brilliant article "The Self and the Future" (Williams 1970). Williams sets up a thought experiment involving a question concerning which of two bodies one will persist in following a certain outlandish procedure. After demonstrating that a compelling case can be made for each of two opposing views, Williams responds to the aporia by admitting "I am not in the least clear which option it would be wise to take if one were presented with them before the experiment. I find that rather disturbing" (ibid., 179).

These explorations have left me in a similar aporetic state. I was a fairly convinced pure contentualist before embarking on these musings, and I have come face to face

with some seriously counterintuitive consequences of my view. I'm considering jumping ship to the dirty contentualist camp. But for now, just considering. I've not given up quite yet. And I think Dennettians, Brandomians and Ismaelians will face similar pressure. But all the other positions on the table seem to face equally unpalatable consequences as well, just different ones. I'm really not sure what to make of this. And *I* find *that* rather disturbing.

## ACKNOWLEDGEMENTS

This paper benefitted from feedback from a number of very patient readers, including Jenann Ismael, Holly Andersen, Andy Clark, and Dan Dennett. Some of the initial ideas were first presented in my 2004 Tamara Horowitz Memorial Lecture at the University of Pittsburgh, and I am grateful for valuable feedback from audience members at that talk.

## REFERENCES

Block, Ned. 1980. "Inverted Earth." Issue: Action Theory and Philosophy of Mind, *Philosophical Perspectives* 4: 53-79.

Block, Ned. 1980. "Troubles With Functionalism." In *Readings in the Philosophy of Psychology*, Vol. I, edited by Ned Block, 268-305. Cambridge, MA: Harvard University Press.

Brandom, Robert. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58: 7-19.

Cosmelli, Diego, and Evan Thompson. 2010. "Embodiment or Envatment? Reflections on the Bodily Basis of Consciousness." In *Enaction: Towards a New Paradigm for Cognitive Science*, edited by John Stewart, Olivier Gapenne, and Ezequiel A. Di Paolo, 361-386. Cambridge, MA: The MIT Press.

Dennett, Daniel. 1981. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: The MIT Press.

Dennett, Daniel. 1992. "The Self as a Center of Narrative Gravity." In *Self and Consciousness: Multiple Perspectives*, edited by Frank S. Kessel, Pamela M. Cole, Dale L. Johnson, and Milton D. Hakel, 103-115. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Flanagan, Owen. 1992. *Consciousness Reconsidered*. Cambridge, MA: The MIT Press.

Grush, Rick. 2003. "In Defense of Some 'Cartesian' Assumptions Concerning the Brain and Its Operation." *Biology and Philosophy* 18: 53-93.

Ismael, Jenann. 2007. *The Situated Self*. New York: Oxford University Press.

Ismael, Jenann. 2011. "Self-Organization and Self-Governance." *Philosophy of the Social Sciences* 41: 327-351.

Lewis, David. 1980. "Mad Pain and Martian Pain." In *Readings in the Philosophy of Psychology*, Vol. I., edited by Ned Block, 216-222. Cambridge, MA: Harvard University Press.

Lewis, David. 1976. "Survival and Identity." In *The Identities of Persons*, edited by Amélie O. Rorty, 17-40. Berkeley, CA: University of California Press.

Metzinger, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: The MIT Press.

Nelson, R. J. 1975. Behaviorism, Finite Automata, and Stimulus Response Theory. *Theory and Decision* 6: 249-267.

Parfit, Derek. 1971. "Personal Identity." *The Philosophical Review* 80: 3-27

Park, Hyeong-Dong, Stéphanie Correia, Antoine Ducorps, and Catherine Tallon-Baudry. 2014. "Spontaneous Fluctuations in Neural Responses to Heartbeats Predict Visual Detection." *Nature Neuroscience* 17: 612-618.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417-457.

Shoemaker, Sydney. 1970. "Persons and Their Pasts." *American Philosophical Quarterly* 7: 269-285.

Williams, Bernard. 1970. "The Self and the Future." *The Philosophical Review* 79: 161-180.