The Hard Problem of the Self
for NWPC

It's not easy to know what is puzzling philosophers when they discuss the existence of the self. Things can become terminological pretty fast. I wish here to outline at least one problem that seems important and non-trivial.

To find what I'm calling the hard problem of the self, we ought, I think, to go back to Descartes and the response to him by Georg Lichtenberg. One of the upshots of the cogito seems to be that I can come to know, even in grave Cartesian doubt, that there is a thing that thinks—and that this thing is me. Thus the existence of a self seems to be proven by the cogito.

But the story doesn't end here. Just because "I am thinking" is true does not necessarily mean that there is a thing that is thinking. In particular, the "I" in "I am thinking" might be a sort of a non-referential pronoun that plays some other function.[1] According to Lichtenberg, in the context of doubt there is no justification for concluding "I am" from "I am thinking" because there is only justification for saying "There is thinking." Just as during a storm we say "It is lightning" without committing ourselves to something that is doing the lightning, so we should not commit ourselves to a self, or

[1] See (Anscombe 1982) Mach.

an “I” when noting the fact that “There is thinking.”

As it turns out, this self-less view is not plausible. Suppose that our conception of thoughts did not require a thinker.[2] In such a case there must be an impersonal way of reporting thought contents along the lines of “There is a thought x”. Now as long as solipsism isn't a necessary truth, it is possible that there are other thinkers, and it is further possible that their thoughts differ. In our normal way of expressing things, we might say that David thinks correctly “I am feeling no pain” while Jim thinks correctly “I feel nothing but pain.” We furthermore believe that these two claims could both be true. But translated into the impersonal, David would have to be thinking “There is no pain” while Jim thinks “There is nothing but pain”. These statements clearly contradict one another, however, so the translation does not succeed. No translation will succeed, in fact, unless the thought contents are relativized appropriately. That is, David must be saying “There is no pain ‘here’” or something of the like, and the same for Jim. The best move for the Cartesian is to maintain that the only appropriate version of “here” is “I.” The self is the relativization point: selves are the places for thought.

It is this very thin notion of a self that interests me, and taken thinly enough we do have excellent reason to believe there are such things. And so far, there is no commitment to what these things could or could not be be—brains, bundles, substances

[2] My argument here is a version of that offered by(Chisholm 1976), and (Williams 1978). A clear version is in appendix O of (Van Cleve 1999).

or souls. Whatever these thin selves are, however, they must be able to play the role of a "place for thought," in the right sense of place.

II.

Before we can become clear on selves, then, we need to become clear on what sense of "place" is appropriate for thoughts, and we need to decide which thoughts should be in that place. So, for example, it initially seems unlikely that we mean place in the literal, spatial sense, but it's not clear what is left over once this is dismissed. And while it is intuitively clear that we want to include my thoughts and Barack Obama's thoughts in different places, what about my unconscious thoughts and my conscious thoughts? What about my beliefs and my pains?

One tempting idea is to embrace the literal and say that the place is simply the brain. It turns out, after all, that's where the mental action is. On this view, when I say "I am in pain" I am right if and only if the pain state is in my brain. I am correct that I am thinking of both apples and oranges iff the parts of my brain that represent apples and the parts that represent oranges are active.

Appealing as it is, a case can be made that being in the same brain is neither necessary nor sufficient for being in the same mental place, in the relevant sense.

Against Sufficiency: Possession and Split Brains

It certainly seems conceivable that there be a brain that supports two radically different mental lives. The most obvious actual case is that of "Split Brain" patients, or patients who have had a commissurotomy in which their corpus callosum, which binds the hemispheres of the brain together, is severed. After the surgery everything seems to be fine, until Robert Sperry enters with his brilliant selective stimulation experiments. [3]

The subject is asked to fix at a particular point on a screen. Then, very briefly—too briefly to allow for eye movement—the words KEY RING are put on the screen, with KEY to the left of the fixation point and RING to the right. The subject is then asked what he saw. He reports that he saw RING, and will deny that he saw KEY. If asked to reach into a bag with his left hand, and pick out what was named by the word he saw, he will pick out a key—even if there are rings and key rings in the bag. What's more, if at one and the same time he is asked to reach into a bag with keys, rings, and key rings with each hand, not looking, he will pick out a key with the left hand and a ring with the right, leaving the key rings behind.

The neurological story behind this mysterious behavior is, of course, rather straightforward. The information flashed on each side of the focal point goes to different hemispheres and the commissurotomy severs the lines of communication

[3] My description of this comes from (Van Cleve 1999) who takes it from (Marks 1980). See also (Nagel 1971).

between the hemispheres. Still, there is no denying how peculiar the case seems. On at least one obvious interpretation of the data there seem to be two mental "spaces" in one brain. This is at least a challenge to the sufficiency of the brain theory of the self, at least insofar as selves are viewed simply as places for thoughts.

Against Necessity: Extended Minds

Once upon a time I had to remember phone numbers, but now I speak a name into my phone and the phone dials the number. If someone asks me for a friend's number, I am useless without my phone—but with it I can come up with the number pretty quickly. One can easily imagine that at some point we could have a chip inserted in our brains that stored tedious things like phone numbers. Several philosophers, most notably Andy Clark and David Chalmers, have argued that there is no reason in principle not to think of such devices as extensions of our minds.[4]

Without going into the details of their argument, they basically maintain that there is no reason to fetishize brain tissue and organic parts as the means by which information is stored and transmitted. If one accepts a roughly functionalist model, beliefs are beliefs not because they are in brains, but because of the role they play in a cognitive system, and so too with desires and cognitive processes such as adding and

[4] See (Clark and Chalmers 1998)

remembering. Beliefs are in brains, but that's a contingent fact, and states outside of my brain play something like the belief role now. I am pretty dependent on my Iphone to remember all sorts of things—appointments, phone numbers, and the addresses of conference hotels. It the fact that I hold it in my hand a reason to say it is not part of my mind? Would we feel any better if Steve Jobs announced a Brain Dock for Iphones so that my Iphone could reside in my head, and could be operated by thought? More argument is needed, but it seems to me that the reasons for denying that external devices are, or at least could be, legitimate parts of the mind are hard to come by. Thus it seems there could be mental states supported by such machines, outside of the brain, that nonetheless were part of the same mind. Thus, being in the same brain is not necessary for occupying the same mental space.

III.

The brain theory should probably respond that one shouldn't count brains by counting lumps of matter, and one shouldn't limit brains to things composes of organic tissue. "Brain" on this view is a functional term, and something that is functionally integrated so completely with a brain, such as a bit of extra memory in an implanted chip, should be considered part of the brain, and isolated tissue in the head should not be.

These responses on behalf of the brain theory are pretty persuasive. But notice that we have changed our notion of "space for thoughts" from actual 3D space, to something like "information space."

What is required for two mental states to be part of the same information space, and therefore part of the same self? One attractive possibility is that two mental states (processes, capacities?) are in the same informational space iff they bear some cognitive relation R to one another. Several candidates for R include the existence of direct inferential connections between the states, the availability of the states to one act of introspection, co-consciousness, or even some more liberal causal proposal such as the ability of states to give rise to one another in a particularly direct way.

The devil is going to be in the details, but there are obviously going to be some questions. How much functional integration is enough? Your thoughts impact mine causally, so not just any causal relation between thoughts will do. Inferential connectedness is promising, but questions remain as to what constitutes an inferential connection—how direct does the connection have to be? Under normal circumstances, commissurotomy patients show a great degree of functional integration. Should we resist our previous intuition that their thoughts occupy different spaces? If I have an unconscious representation of my sister as an evil person, and this only occasionally manifests itself in a Freudian slip when I call her "Satan" instead of "Sarah", is this

connected enough? One suspects that vagueness abounds hereabouts and that there will not be one answer to whether or not two thoughts occupy the same space.[5]

As long as we are focused on functional integration, or some sort of inferential or computational connectedness, I think it is likely that we are going to run into selves with vague boundaries. Or, to put it another way, it will be indeterminate whether or not two states are states of the same self. Just as it is probably indeterminate whether a bunch of processors and hard drives consitute one computer as opposed to several, so it will be indeterminate what constitutes one versus many selves.

This seems worrying not only because metaphysical vaguess is unattractive, but because selves just don't seem like the sorts of things to be vague. How could it be indeterminate whether it was me thinking a thought? Could it really be a contextual matter, as it is in other cases of vagueness and indeterminacy?

Perhaps there is some context in which Jim and David work so closely together that, like a pair dressed in a horse suit, they can sometimes be described as one person. It seems like this is a very attenuated, artificial sense, and that there is a more fundamental sense—not just different, but more fundamental—in which David is quite correct to say "I don't have a headache" in cases when Jim does have a headache.

How do we reconcile the Cartesian intuition about the existence of a single, non-

[5] (Nagel 1971)

vague place for thoughts with the fact that spaces for thoughts seem to be functionally defined, admitting of degrees of integration? The Cartesian should probably insist that defining selves in terms of functional integration is not the way to go, and that when you leave the occurrently conscious, or even the phenomenally conscious states, one is letting vagueness in the door, but not earlier. There is a relation R, that does not come in degrees, and that is not vague—the relation two conscious states have to one another when they are co-conscious. The asymmetry between this relation and the others we have discussed marks a metaphysical difference between them, and this justifies our saying that there is in fact a non-arbitrary notion of the self which is fundamental, and which is ultimately at the heart of any other notions of self(Dainton 2008)[6]

This approach would tie the Cartesian intuitions about the self quite closely to the issue of the unity of consciousness. In particular, it would tie them to what Chalmers and Bayne call the "phenomenal unity of consciousness." According to this thesis:

> **Phenomenal Unity Thesis**: Necessarily, any set of phenomenal states of a subject at a time is phenomenally unified.[7]

Phenomenal unity is defined as follows:

> [A] set of conscious states is phenomenally unified if there is something it is like

[6] (Dainton 2000)
[7] (Bayne and Chalmers 2003) p.13

for a subject to have all the members of the set at once, and if this phenomenology subsumes the phenomenology of the individual states.[8]

So what makes it true to say there are two selves in the Jim and Dave horse is that the pain and the conscious thought there is no pain, say, are not co-conscious. There is no one state that includes both of them such that there is something that it is like to have that state, and that if one has that state one also has each of the states that make it up. This suggests the following notion of the self:

Phenomenal Self: The self is the space of co-conscious phenomenal states. Phenomenal states that are co-conscious are states of a single self, and phenomenal states that are not co-conscious are states of different selves.

To me this view has some plausibility, in part because we are directly aware, even acquainted, one might say, with the states that are supposed to be in the relation at hand, and it seems difficult to imagine counterexamples to the phenomenal unity thesis.

It is difficult to know what exactly it is like to be a commissurotomy patient. It could of course be that when it comes to the information in their right hemispheres they are like super blind-sighters—they have access to the information, but there is no phenomenal consciousness that accompanies it. They do not have phenomenally conscious states that correspond to what is flashed to the left part of their visual fields.

[8] (Bayne and Chalmers 2003)

If that is how it is, then this phenomenal view would hold there is only one self, with unconscious information controlling behavior elsewhere. It is perhaps more natural, though, to guess that there is phenomenal consciousness that accompanies both pieces of information, but that those phenomenal states are not co-conscious, or, to put it another way, that there is not a single phenomenal state of which they are both a part. In this case, it would be correct to say that there are really two selves there in this sense.

The extended mind cases do not obviously present any trouble for this either. Whether my Iphone is part of my mind or not, it is certainly not a source of occurrent psychological states or phenomenal states. But, if in the future there is a way to extend the parts of the brain that underlie consciousness, there is still nothing in principle problematic. If the states in this outer device are co-conscious with all of the other states, there is one self there. Otherwise, not.

IV.

The title of this paper is an homage to what David Chalmers called "The Hard Problem of Consciousness."[9] At this point the connection to the consciousness puzzle is pretty clear. In both the case of consciousness and the self, if one focuses upon cognitive states alone, it is hard to get at the puzzle. The puzzles in these areas only

[9] (Chalmers 1996)

really become pressing from the point of view of the subject—from the bearer of consciousness. There is, in other words, a sense in which we have a difficult time finding the self in an objective theory—that is, a theory that can be fully apprehended without occupying any particular point of view. From the outside, the world does not seem to break up into discrete loci of consciousness because from the outside the relation of co-consciousness does not seem to appear as a distinct, or at least an important, relation at all. This is not to say, of course, that there are not neural relations that are responsible for or even identical to co-consciousness. I am not making that metaphysical step. I only want to argue that if one finds there to be an explanatory gap in the case of consciousness, which is seems the majority of philosophers do, then one should perhaps find the same gap in the case of the self. This makes sense, I think, of some of the odd claims one sometimes finds in Sartre, Kant, and others in the post-Hegelian tradition: that the subject is essentially subjective and cannot be objectified. It also makes sense, I think, of some of the intuitions behind substance dualism. If there is an explanatory gap for the self, the epistemological step for arguments for dualism is in place. I don't mean to endorse these arguments or claims, but their connection to the problem of the self makes it clear that the problem is important, non-trivial, and worth the attention of contemporary philosophers.

Anscombe, G.E.M. 1982. The First Person. In *Metaphysics and the Philosophy of Mind: Collected Philosophical Papers Vol. II*. Oxford: Basil Blackwell.

Bayne, Timothy J., and David J. Chalmers. 2003. What is the unity of consciousness? In *The Unity of Consciousness*, edited by A. Cleeremans: Oxford University Press.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*: Oxford University Press.

Chisholm, Roderick M. 1976. *Person and Object*. LaSalle: Open Court.

Clark, Andy, and David J. Chalmers. 1998. The extended mind. *Analysis* 58 (1):7-19.

Dainton, Barry. 2008. *The Phenomenal Self*: Oxford University Press.

Dainton, Barry F. 2000. *Stream of Consciousness: Unity and Continuity in Conscious Experience*: Routledge.

Marks, Charles. 1980. *Commissurotomy, Consciousness, and the Unity of Mind*. Montgomery: Bradford Books.

Nagel, Thomas. 1971. Brain Bissection and the Unity of Consciousness. *Synthese* 22:396-413.

Van Cleve, James. 1999. *Problems from Kant*. New York: Oxford.

Williams, Bernard. 1978. *Descartes: The Project of Pure Enquiry*. London: Penguin.