Addleton Academic Publishers
Linguistic and Philosophical Investigations
Volume 12, 2013, pp. 47-66, ISSN 1841-2394

THE EXTENDED SELF, FUNCTIONAL CONSTANCY, AND PERSONAL IDENTITY

JOSHUA FOST
jwfost@pdx.edu
Philosophy Department
Portland State University

ABSTRACT. Personal indexicals are often taken to refer to the agent of an expression's context, but deviant uses (e.g. 'I'm parked out back') complicate matters. I argue that personal indexicals refer to the extended self of the agent, where the extended self is a mereological chimera incorporating whatever determines our behavioral capacities. To ascertain the persistence conditions of personal identity, I propose a method for selecting a level of description and a set of functional properties at that level that remain constant over a lifetime. I argue for functional constancy, and against continuity, as the central determinant of diachronic identity.

Keywords: Functionalism; Extended mind thesis; Personal indexicals; Personal identity; Psychological continuity theory

1. Introduction

There are two investigations of personal identity that need interrelating. The first is the linguistic study of what personal indexicals ('I', 'me', etc.) refer to; that study's goal is a theory that systematizes as many uses of these terms as possible. The default view is that 'I' just refers to the utterer, but that can't be the whole story: witness deviant uses like 'I'm parked out back.' Such uses suggest that 'I' can refer to something other than a mind-body conjunct—i.e. the default composition of the self. If that's right, then semantic theories that ground personal indexicals in the self have, at best, passed the buck, not clarified the de facto meanings of these terms. The second investigation, to which the first relates closely, is the effort to discern a reasonable set of conditions for how selves are individuated and how they persist over time. Linguistic clues betray assumptions of both unity and persistence: 'I refers to *the* utterer.' This second investigation encounters its own set of obstacles, like how to deal with compositional (e.g. somatic) changes, smooth functional variation, psychological discontinuities, and more.

My purpose in this paper is to tie these two investigations together by probing the implicit philosophy of identity that hides behind some of the things people say. Most of the recent literature on personal identity considers rather exotic thought experiments involving brain transplants and the like in order to—supposedly—shed light on our intuitions. I will depart from this practice and consider more everyday scenarios because I don't think that such scenarios figure in to the common understanding of personal identity.

As a disclosure of my overall stance, let me say that I agree with Derek Parfit (1995) that many of the problem cases surrounding personal identity, including some of the cases I discuss here, are merely decision problems in the sense that their primary impact is in exposing personal judgments about whether two entities (at different times, let us suppose) deserve to be regarded as the same person. Here's Parfit:

> Some conceptual questions are well worth discussing. But questions about personal identity…are like questions that we would all think trivial. It is quite uninteresting whether, with half its components replaced, I still have the same audio system. In the same way, we should regard it as quite uninteresting whether, if half of my body were simultaneously replaced, I would still exist. As questions about reality, these are entirely empty. Nor, as conceptual questions, do they need answers. We might need, for legal purposes, to give such questions answers. Thus we might decide that an audio system should be called the same if its new components cost less than half its original price. And we might decide to say that I would continue to exist as long as less than half my body were replaced. But these are not answers to conceptual questions; they are mere decisions. (Parfit, 1995, p. 302)

On this view, most of the discussion around personal identity just assumes that personal identity is a natural kind, and that our philosophical project is properly taken to be figuring out what that kind consists in. I reject both assumptions: I just don't think there is any reason to regard personal identity as 'being' anything other than an idiosyncratic concept whose rough-and-ready machinery is practicable for many ordinary uses but which breaks down in some mildly unusual problem cases like persistent vegetative states, and some downright exotic ones, like tele-transporters, fission, and brain-state transfers. This is not at all to say, however, that this popular philosophical project is a waste of time: by exploring the structure of concepts—as exposed by language—we reveal the consistency, entailments, and groundedness of our models of the world and of ourselves, and I think that is a perfectly reasonable undertaking for philosophers (and cognitive scientists).

I will defend two main claims. These will be that (1) our de facto use of personal indexicals implies that we normally think of personal identity as being *extended* in the sense of Clark & Chalmers (1998), and (2) functional *constancy*, not continuity, provides the better foundation for diachronic self-identity, despite there being no normative criterion for the functional traits that matter.

The structure of the paper will be as follows: In the first section, I'll review Kaplan's (1989) account of personal indexicals, present some challenges to it, and end up defending it in a form modified to include the extended self. After that, I'll discuss the ways I think functionalism can and should bear on this inquiry. That will lead to a discussion of Shoemaker's psychological continuity thesis and Nozick's theory of the closest continuer. I'll argue against both of those positions and describe how the functions that constitute selves can remain constant, and thereby form the basis for diachronic persistence.

2. Kaplan's Model, Deviant Uses, and the Extended Self

Some words seem to refer to the same thing no matter who says them: 'Mt. Everest' always refers to the same mountain. Other words, like 'I,' as in 'I climbed Mt. Everest,' mean different things at different times: it means Mary when she says it and it means John when *he* says it. That shifting meaning is what makes a term an indexical. Kaplan's (1989) systematizes this by asserting that linguistic expressions have *contexts,* where a context is a tuple <w, a, p, t>, and *w* is a world, *a* an agent, *p* a place, and *t* a time. Some indexicals, which Kaplan calls *true demonstratives,* use extra-linguistic behavior from the speaker to fix their meaning, as when a speaker, pointing to a donut, says 'I want *that*.' In contrast, *pure indexical* terms like 'I' require no special behavior from the speaker. Personal indexicals are exemplars of this category: they always refer to the agent of the context and therefore to the whomever the utterer happens to be.

Kaplan's semantics, then, have foundations based on the agent. One of my concerns is that unless we understand what agents are—or, more specifically, unless we understand how they are individuated—Kaplan's model isn't enough to fully understand what 'I' means. Why would we be confused about how agents are individuated? Change over time, for one. As I'll discuss in detail later, expressions like 'I'm not the man I once was' suggest *two* identities, one former and one present. More exotically, split brain personality and multiple personality disorder challenge the unitariness of agents. Or suppose we started removing neurons from a person's brain, one by one. Eventually we would find ourselves empty-handed: when did the collection of remaining cells stop being *the agent*? (This particular example has the disquieting property of having practical application, as when brain cells slowly die off. I'll be discussing this in some detail later.) What happens when the psycho-functional properties of an agent vary through learning? Absent answers to such questions, Kaplan's grounding on *the agent* could refer to any number of different things. Which one?

Before I try to answer that, I want to explore a bit more thoroughly what ought to be a simpler question. Suppose we just posit that we know what agents are—suppose people never changed, never had multiple personality disorder, etc. Under those simplified conditions, would 'I' always refer to the agent? If so, then the

agent is stranger than is typically supposed, as shown by non-canonical uses of 'I' in locutions like the following (from Mount, 2008 and Krasner, 2006):

- 'I'm parked out back' (where 'I' here refers to my car);
- 'I am on a purple square' (where 'I' refers to a player's Monopoly game piece);
- 'I killed a woman yesterday. Naturally I want to deflect suspicion,' (where the speaker here is a detective trying to put himself into the mind of a murderer, and 'I' refers to that person).

On their face, these seem to trouble Kaplan's model because in each case 'I' fails to refer to (what is typically regarded as) the agent of the utterance.

A conventional response is to distinguish between semantics and pragmatics. Kaplan's model concerns the former, and is aimed at systematizing the formal structures of linguistic reference. The latter concerns the ways in which competent speakers use a language to communicate. Theories of implicature in particular are concerned with explaining how speakers communicate meanings that differ from the literal content of their utterances (as, for example, in using the phrase, 'Yeah, right!' to mean 'Absolutely not.'). Deviant uses of 'I,' on this view, would fall under the umbrella of implicature.

This solution makes me uneasy. It sets a precedent of abdication, wherein semantic theories push off onto pragmatics whatever they can't explain. Suppose, therefore, that one thinks that deviant uses need to be included semantics per se. Mount (2008) thinks that they do. The examples above, on her view, are not social conventions or illustrations of artful expression, but straightforward applications of 'I' and therefore reflections of its formal role in language proper. Her proposed solution is to include the *intentions of the utterer* in semantics itself. This allows a shifting of reference, even for pure indexicals, that stands in clear contrast to Kaplan's theory, where 'I' has a fully automatic meaning in every expression. Here is Mount's summary:

> I have argued that the automatic indexical theory fails to capture how indexical reference works, and that discretionary speaker intentions can affect the reference of all indexicals in some way or other. In other words, there are no 'pure' indexicals. This conclusion is compatible with various conceptions of how indexical reference works, but it is not compatible with any theory according to which indexicals refer automatically and invariably based on a certain specifiable feature of the objective utterance situation.…[S]peaker intentions play a role in the *meaning* of indexicals. While I don't think that speaker intentions themselves *determine* what an indexical refers to, they do have a predictable sort of influence. (p. 208) [Emphasis is original.]

Thus, on Mount's view, 'I' is a *discretionary* indexical lacking invariant reference, in contrast with Kaplan's theory, which makes 'I' an *automatic* or *pure* indexical that invariantly refers to the agent of a context. In the next few paragraphs I'll

formulate my response to Mount and defend a modified form of Kaplan's theory. The first part of my response rests on our tendency to construct an 'extended self' in our cognitive representations, so my immediate purpose is to talk about what that means.

The *extended mind thesis* was proposed by Clark & Chalmers (1998) as a way to accommodate our use of external artifacts as cognitive accessories. Using their example: if Otto, who has Alzheimer's disease, uses a notebook to store the address of a museum which he could not otherwise remember, the typical view is that Otto knows that he can use the notebook to retrieve that information. Without such use, Otto himself does not know the information. The extended mind thesis, on the contrary, says that Otto does know the information—it's just that some of the information Otto knows is contained not in his brain, but in the notebook. The corollary *extended self thesis* (EST) goes one step further, saying that the extended mind thesis pushes us to conclude that *Otto's self includes the notebook*, not just the biological material comprising his animal body. Our selves spread out, beyond our skin. I am going to take *the self* to be the same thing as *the agent*. Agents, then, spread out beyond the skin.

Now I want to use the EST to address the problem cases discussed by Mount. One of these is the parking example, as in 'I'm parked out back' or even 'I'm out back.' The EST can clarify what's happening here by asserting that the car, like the notebook, is a bearer of self-extension. That is, the person making this utterance has a conception of the self that includes not just their own mind and body, but their car as well. (I am well aware that this rather a different sort of extension than Otto and his notebook; I'll address this different use in a moment.) Consider some related utterances that seem to support the assertion that people think of their cars as part of their selves. Fender-bender victims says things like 'Hey, he hit me!' When parallel parking, I ask my passenger, 'Can I fit in that space?' Damage to the car I am driving (think of scraping its underside on a parking ramp) elicits wincing and other expressions of pain, even though no damage is done to my animal body. These examples reflect the degree to which physical proxies, especially those whose movements we can control precisely and whose physical interactions have consequences for us, are regarded by us *as* parts of us. As reflections of our cognition, then, they underwrite utterances like, 'I am out back' as meaning 'I (the mereological sum of my body and my car) am out back.'

These examples offer circumstantial support for my claim that the car acts as a bearer of self-extension, but some may object that this application of the EST overextends it by including an external object that, unlike Otto's information-containing notebook, does not participate in cognition. This objection rests on an unnecessarily narrow view of the EST. The central feature of the EST is the posit that a coupling between two or more systems can produce a new entity that can be regarded as a (composite, but so are 'merely animal' human beings) system in its own right, with its own new behavioral capacities. In the Otto example, the

information in the notebook is what provides those capacities. The notebook encodes information which, when coupled to the rest of Otto, comprises a composite system with behavioral capacities that neither the notebook nor Otto-minus-the-notebook has alone. The encoded information is not an abstraction: it is embodied in the arrangement of ink molecules on the notebook's pages. We are really dealing here with a coupling of two physical systems. Although the car does not encode information in the same way, this difference does not bear upon its satisfaction of the central feature of the EST. It (the car) extends the behavioral capacities of a driver just when the two interact in certain ways. To accommodate and make use of this behavioral extension, the composite system is represented as an extended self by the self-representing parts of the system (i.e. the brain).

All of this goes to show that Mount's first example 'I'm parked out back' / 'I'm out back' does not require speaker intentions to fix the meaning of personal indexicals. They can still refer to the agent, so long as the agent is understood under the umbrella of the extended self thesis.

Mount's second example is Monopoly, as in the utterance 'I am on a purple square.' The EST can be applied here by noting the coupling between the player and the game piece. The player can move the game piece, but, per the rules of the game, there are constraints on these movements. Certain fates for the game piece, like landing on a purple square, compel certain behaviors on the part of the player. This is very much like the coupling between car and driver, and satisfies the structural conditions of the EST. The end result for cognition is that the game piece, like the car, becomes, in the player's mind, a part of their extended self. All the attendant statements then make sense: 'I am the little dog,' 'Please don't put me in jail,' etc.

Now, semantically, this kind of extension might cause a problem: How is a listener to know which *part* of the extended self I mean when I use personal indexicals? Consider Mount's example (p. 208):

Player 1: Hooray! I'm finally on a purple square.
Player 2: Yeah, but the blue square I'm on is better.
Player 1: I've had an annoying song stuck in my head all day. [Rolling the die] Good, now I'm on another purple square.

Mount thinks, again, that this under-determination requires speaker intentions to fix the reference. In other words, she thinks that in Player 1's first utterance, 'I' refers to the game piece, while in first part of Player 1's second utterance ('I've had an annoying song stuck in my head'), 'I' refers to the (non-extended) human animal player. Finally the reference goes back to the game piece. To resolve this ambiguity, says Mount, requires the additional switching mechanism of speaker intention. Those intentions are received somehow by hearers, and the combined effort resolves the ambiguity.

I think Mount is probably right that this effortful perspective-sharing often plays a critical role in allowing hearers to discern speakers' intentions, but I don't think that this applies more forcefully to personal indexicals than other terms. At least, we do not need to assume that it does, because vagueness concerning which part of a multi-part entity is meant is common in references even to the *non*-extended self: 'I hurt my hand' needs something more to fix which part of the hand is meant, but we don't wonder about the meaning of 'hand' or say that something special is required to determine what that term refers to. 'Hand' is a vague term and so is 'I'. My account of the Player1 / Player 2 example above, therefore, is that 'I' refers to the multi-part human-animal-plus-game-piece entity, i.e. the extended self of the agent. Under-determination is unproblematic because the content of each statement unambiguously picks out which part of the extended self is meant, even *without* effortful perspective-sharing. Unless the speakers themselves are sitting on a floor with colored squares, Player 1's first utterance ('I'm on a purple square') can only refer to the game piece part of the extended self, while the first part of Player 1's second utterance ('I've had an annoying song…') can only refer to the human animal part. Moreover, in the earlier parking example, 'I'm out back' is literally true under the auspices of the EST: As a mereological sum, I can be both here and out back at the same time. In none of these cases does the reference actually shift: 'I' always refers to the extended self, and rest of the content of the utterance clarifies—to the extent needed for pragmatic communication—which part of that self is meant. Therefore we don't need Mount's speaker intention or hearer effort to fix the meaning. If, on the other hand, 'is on a purple square' could have been reasonably predicated of both the human animal *and* the game piece, then something more than the automatic meaning would be required. Even then, however, the semantic situation would be no more confusing than in statements having nothing to do with personal indexicals, and therefore outside the purview of the EST and my proposal.

What about the last example with the detective ('I killed a woman yesterday…')? One account of this expression, which I like but Mount does not, is that it is pretense. The speaker is pretending to be someone else. This pretense will be clear in the context of a detective speaking to a colleague, so that the colleague will understand what the detective means by 'I'. Mount dislikes this account because it lacks scope by not applying in other situations, including the car example (no one *pretends* to be a car), and presumably she is hoping for a semantic theory that encompasses all uses. My response to this is that, first, Mount's own proposal of speaker intention does not provide a theory that encompasses all uses. It is merely a wildcard that allows personal indexicals to mean all sorts of different things. Thus her own account does no better with respect to explanatory scope. Second, I just don't see why Kaplan's model doesn't work here. Yes, the detective is pretending to be the killer. When he speaks, 'I' refers to the agent, as always—but the agent, under the pretense, *is* the killer. For that matter, the detective could have said, 'I'm parked out back for privacy while

moving the body,' thus referring to the part of himself that is the killer's car, per the EST.

3. Functionalism and the Constancy of the Agent

So far I have been focusing on deviant uses and extension, ignoring the important question about whether and/or how selves can change over time. It is to that question that I turn now.

By themselves, neither the extended self thesis nor Kaplan's model say anything about the persistence of agents. They would both apply equally well to a world so volatile that agents flickered in and out of existence, reborn with different properties each time. In such a world, there would still be linguistic contexts, and agents could still utter things. But that possible world is not our actual world. When I say 'I read the newspaper yesterday,' I mean that the agent uttering the sentence is the same agent as the one who read the newspaper. On the other hand, the idiom 'I am not the man I once was' suggests a break: the agent uttering the idiom is asserted to be, in some important way, a different agent than the one from the past. Indeed, the signaling of a deviation from the default assumption of agential constancy is the canonical purpose of the idiom: it implies that something unusual has happened, that I have changed in ways significant enough so as to interrupt the default diachronic constancy of my identity.

If people *think* that their selves are ordinarily constant over time, then that is what they mean when they refer to themselves, and our semantic theory of personal indexicals ought to capture that meaning. This brings me to the next phase of my argument. To this point, I have augmented Kaplan's theory by suggesting that personal indexicals refer to the extended self of the agent of a context. My next step is to reframe the extended self itself so that it can persist in the way we need it to for our expressions to carry our intended meaning.

The biological variability of our bodies and the psychological variability of our minds ought to be enough to give serious pause to proposals that a persistent personal identity consists just in the constant composition of the body or the function of the mind, and these have been important motivations for proposals that focus on (a) (usually non-branching) continuity rather than constancy (Shoemaker, 1984; Parfit, 1984); (b) causal inheritance; and (c) sufficient but not necessary conditions (Davis, 1998). Bucking this trend, I'll be arguing that there *is* something necessarily constant about selves, and that that something grounds our notions of agential persistence. Given our tolerance for morphological changes (I can lose all my limbs, triple my weight, and get a face transplant and still be regarded by most people as still me), I think it's obvious that something about the mind is the de facto basis for commonsense views of self persistence. The abstract nature of that something, and additional concerns like the desire to avoid substrate

chauvinism, leads quickly to functionalism, so I should say what I mean by that term.

By *functionalism* I mean the thesis that minds are fully characterized by the set of causal relations between mental states. A mental state, in turn, is a vector, each component of which is a degree of freedom for the mind. Such vector spaces can be constructed at many levels of description—a degree of freedom could be as concrete as a neuronal firing rate or as abstract as the level of emotional arousal—a fact I'll be making use of below.

Under this definition, functionalism is a thesis just about minds, but it can easily be generalized to work for other phenomena that we normally construe as discrete and durable despite their supervenience on volatile substrates. The more general functionalist thesis could then be framed as the assertion that entities are fully characterized by the causal relations between their states. A Leibniz's Law-like version would look something like this:

$$(x)(y)[(x = y) \leftrightarrow (C)(i)(j)(x_i C x_j \leftrightarrow y_i C y_j)]$$

where C is a causal relation, x and y are entities, and $x_i \in X$ is a state of x (similarly for y and j). It will be easiest just to think of C as deterministic causation; in that case two things are identical just when their state spaces are the same and they always flow the same way through them.

With that foundation, denote by F the collection of all the inter-state causal relations for a particular entity, i.e. the function that maps $x_i \rightarrow x_j$. Together, F and X define a *type* of thing; any entity satisfying the relations of F is a token of that type. Note that if two entities are identical under this criterion, their overall input/output relation with the world will be identical because their identical flow through state space will necessarily result in their identical reaction to a given set of external circumstances.

Applying this to the present context of personal identity yields something like the following. First we define the state space. This just means listing out all the degrees of freedom for a mind, at some level of granularity. In common practice, this will include many of the posits of folk psychology: levels of belief in propositions, levels of hedonic value assigned to different experiences, etc. Then we define F by describing the way that mental states evolve over time. This will include mundane inter-state relations like eating when hungry to specific rules that uniquely characterize individuals: when I see *this* photograph I am put into such-and-such mnemonic / emotional state. To tell what constitutes my personal identity—to say what makes me me—I pick out the portions of F which, in my judgment, matter the most.

Suppose I have done all that. If I then say, 'I read the newspaper yesterday,' what I mean is that the functions that I regard as *constituting* myself—the part of F that I picked out as mattering—has not changed since yesterday. I am aware that some other parts of F *have* changed. I might have (knowingly or not) a slightly worse recall for some tidbit of current events. By saying that I am the same person who read the newspaper yesterday while acknowledging that my memory for this

tidbit has faded, I assert implicitly that that memory is not one of the psychological features that defines me. The same goes for many other incidental changes. My opinion of another person has changed slightly since yesterday. That opinion, though it is a functional relation between mental states (specifically, it is a relation between (a) the state of exercising my concept of the person and (b) a certain judgmental or emotional state) does not belong to the set of relations that I take to define me.

When I say 'I am not the man I once was,' something in the idiomatic subtext suggests that I am referring to a functional relation that *is* important enough to me so as to constitute a change in agent. This would typically be taken to be a personality-level change: a core value, motivation, broad-based behavioral disposition, etc. Inasmuch as some definitional set of those relations held once but no longer, I am not the same man.

It will always be possible to choose some set of functions—some parts of F—that remain constant over a lifetime despite lower level variability. That set of functions, says the functionalist, can serve as the referent of personal indexicals and thereby satisfy the persistence conditions that I usually endorse implicitly when I talk about myself.

Let me say this one last way. I can describe my mind at many levels of abstraction. With relatively low abstraction, I can talk about my brain's sub-neuronal constitution. This constitution gives rise to particular functional properties for my mind. At that level, some functional properties will be difficult or impossible to pin down precisely: I probably cannot give a spike-for-spike account of how an individual neuron will react to a given stimulus and expect that relation to hold for many years. At higher levels of abstraction, though, I probably *can* give functions that are constant over spans of decades—for example, a generally pleasurable response to certain types of natural scenes. It is in sets of higher-level functional abstractions like this that we can find a self that is constant over time.

This clearly means that there could be, in principle, multiple instances of the same mind. This implication is no more troubling, I think, than the uncontroversial idea that there can be multiple instances of the same book. At least, that is uncontroversial when books are defined in terms of their informational content. A book defined more deeply, e.g. in terms of its physical constitution, is soon embroiled in the same identity and persistence questions as those found in the philosophical discussion of persons and selves. It is an accidental feature of the world that, in the way we typically define them, minds are *sui generis*: created by processes that produce just a single token, while books and electrons are not. There is nothing strange in saying 'I have ten copies of the same book'; it is mere happenstance that we are less comfortable with 'My brother is in ten countries right now.' (Incidentally, that expression would make conventional sense if the brother were the host of a syndicated radio show; but in that case, the semantics

suggest that the referent of 'my brother' includes the radio show—a use sanctioned by my application of the extended self thesis.)

Another potential concern with my constancy account is that we might be mistaken about what has remained constant about us, and define F so that there is no entity satisfying it. We might, so this concern goes, be told by a diligent third-party investigator that, despite our intuitions about the stability of our personal identity, despite referring to and thinking about ourselves constantly, the term 'I' failed to refer to any actual person.

This concern can be addressed in several ways. The simplest—and not very interesting way—is to note that many terms refer to concepts with no embodied instances in the real world. Pegasus, for instance. One might erroneously believe that such terms did refer. Similarly we might have a concept of a person we thought we ourselves embodied, but be mistaken about this.

A second, more interesting response, is to note that there will be *some* set of functions, some subset of F, that remains constant over our lifetime. These constancies may not be fine-grained features of personality but, perhaps, more generic features of human psychology. Perhaps severe amnesia has taken our memory, dementia our reason, and trauma our emotional, gestural, non-verbal habits. There will still be more general functional dispositions that remain constant, such as the tendency to startle to a loud sound, or to feel relief when that sound was revealed to be unthreatening. For a person unfortunate enough to have suffered so many losses, these generic functions will indeed be the only self that has survived their entire life. For more typical people, smaller changes—events vivid in one's teenage years will be entirely forgotten a few decades later—do indeed mean that the teenage self is not the same as the adult self, but they don't mean that *no* self has persisted. (It should be obvious that possible changes are not just losses. As I age, I gain expertise, acquire new values, etc. and these functional additions make me someone different from who I was when younger.)

This position, in other words, does not guarantee the persistence of a *given* agent. It *does* guarantee the persistence of *some* agent: we can always find a set of functions that remains constant. Going the other way—defining a particular agent as being comprised of a particular set of functions—leaves open the possibility that lower-level volatility will undermine the implementation of those functions and thereby extinguish the person who was defined by them.

I cannot resist bringing in here another line of inquiry. It starts with the sorites I mentioned earlier: suppose my brain perishes, cell by cell. At first I am clearly myself. At the end, there is an empty skull that is pretty clearly not me. Yet intuitively a self constituted by n neurons is the same self as the one constituted by $n-1$ neurons.

As it happens, the progression of Alzheimer's Disease (AD) is not far removed from this scenario, and we can therefore ask the tragic but informative question: What happens to a person's conception of *their own* personal identity as their brain gradually disappears? Tappen et al. (1999) analyzed conversations with mid- and

late-term AD patients, recording the appropriate and inappropriate use of personal indexicals. They concluded that even in this population, personal indexicals were used 'frequently, freely, and coherently' (p. 5). This is despite the fact that these same patients unequivocally suffered from memory loss, deterioration of cognition and linguistic ability, emotional changes, and the widespread perception on the part of friends and family that the patient's *self* was 'gradually fading away' (Tappen et al., p. 2). The patients themselves were often aware that something was wrong with them, that something about their cognitive skills and memory had changed. Note that this experiment probed the proper linguistic use of the terms: it did not examine the depth of those terms' referents. It could be, that is, that the patients' coherent use of 'I' and 'my' and so on reflected not the stability of what they meant by those terms, but the inevitable presence of a 'center of narrative gravity' (Dennett, 1992) as long as narration was possible. Even as such a linguistic entity persists, the full texture of its referent could be dissolving. This possibility is supported by Addis & Tippett (2004), who explored the impact on identity of autobiographical memory loss in AD patients. They found, in brief, that several measures of identity, including strength (as measured by the number of answers given to prompts of 'I am…') and quality (as measured by the specificity of those answers) were reduced in AD patients, and that these reductions were correlated with the loss of personal autobiographical memories.

Perhaps these experiments reveal something about conceptual role semantics: if the meaning of 'I' is the *content* of a personal identity, i.e. the collection of psychological features that remain constant, then as that collection deteriorates, one might think that speakers' facility with 'I' ought to deteriorate correspondingly. This did not happen in Tappen's sample. There, speakers' semantically proper use of 'I' remained undisturbed despite the deterioration of what had once been psychological constancies. On the other hand, perhaps the patient's awareness that something was wrong would lead them to say that they were no longer the same person. I'm not aware of any studies addressing that particular question. Finally, it's possible that the patients' loss of autobiographical memory prevents them from realizing the full extent of what has changed, and therefore that they, and we, cannot make well-informed judgments about whether their selves have persisted through the turbulent course of the disease.

My account of constancy is open, of course, to the possibility that the functions that remain constant—and therefore constitute the identity of—person A are different from the functions that constitute the identity of person B. It may be, in other words, that Alice's interest in novel tastes has remained constant while Bob's has varied. Meanwhile, Bob's frustration with injustice could have remained constant while Alice's varied. In developing the 'closest continuer' theory of personal identity (more on this below), Nozick (2003) makes the same point:

> I do not believe that there are fixed details to be filled in; I do not believe there is some one metric space in which to measure closeness

> for each of our identities. The content of the measure of closeness, and so the content of a person's identity through time, can vary (somewhat) from person to person. What is special about people, about selves, is that what constitutes their identity through time is partially determined by their own conception of themselves, a conception which may vary, perhaps appropriately does vary, from person to person. (p. 113)

This also echoes Parfit's views, quoted earlier, about the non-normativity of identity of an audio system. It's important to note, though, that not everything is relative. Once we've decided upon a function, we can be quite rigorous about whether it has or hasn't remained constant. At the microphysical level, the properties determining the functional behavior of particles are well-understood: mass, charge, spin, etc. If any of those properties change, our broader theory of physics will tell us the functional difference that results. At the psychological level it is possible to quantify at least some of the properties determining input/output relations: the probability of accepting an apple pie when hungry, the level of stated agreement with some philosophical theory, the percent accuracy of autobiographical memories of the last World Series. If the psychological supervenes on the physical, then it is possible *in principle* to know whether, despite cellular turnover, various synaptic changes, etc., Alice's interest in novel tastes has remained constant. In practice, of course, this will not be possible. In practice, all we can do to determine whether the Alice of today is the same as the Alice of a decade ago is probe a few salient features and make our guess. The same goes, for that matter, for Alice herself.

Relating this all to the extended self is easy. While the psychological version of functionalism limits the state space X to the mental, generalized functionalism imposes no such constraint. Components of the self that reach out beyond the skin can also be described in terms of state spaces and causal flows; in the method I'm adumbrating here, those degrees of freedom and causal relations are just grafted to the psychological model to produce a functional description of a chimeric (i.e. extended) entity.

Applying this to the original example of Otto and his notebook: the information in the notebook is asserted (by proponents of the EST) to be a part of the total information in Otto's mind. Very crudely, the state space of the notebook could be *n*-dimensional, with one dimension for each written character. Since the notebook does not do much of interest by itself, we might just neglect the inter-state transitions for it and think only of the way Otto's mind interacts with it. But if the information in the notebook is critical to his self-identity (as judged by Otto or anyone else who wants to know), then its erasure should erode Otto's identity in the same way as would a brain injury in a non-extended self. Clark and Chalmers imagined the notebook to contain the address of a museum, but it might have included much more than that—autobiographical memories, for instance, of whatever scope. If that information is stored only in the notebook, but is used by Otto to maintain modes of behavior that uniquely characterize him, then I think it

would be reasonable to say that the loss of the notebook—and the concomitant excision of an important part of F—would result in Otto's functional death. *Someone* would persist, but it would not be the same person. Looking at the sorts of things real people say when they lose information storage and processing devices (cellphones, personal organizers, web access, etc.) lends anecdotal support this claim, and helps measure the impact of amputations to the extended self.

4. The Continuity Requirement

There are theories of personal identity that have nothing to do with functionalism; space prevents me from addressing them here. What I do want to address here are functionalist theories that do not rest on *constancy*, but something else. The pivotal one is the theory created by Shoemaker (1984, 2004) and Parfit (1984) which rests psychological *continuity*. A second one, really a refinement of the first, is Nozick's (2003) *closest continuer* theory. I'll address each in turn.

In one of its simpler forms, the Shoemaker/Parfit idea—call it the *psychological continuity thesis* (PCT)—is that iff a psyche (composed of memories, beliefs, desires, etc.) at time t_2 is causally inherited in certain ways from a psyche at time t_1, then the two psyches belong to the same person. This bears on the account I'm giving in this paper because psychological variation per se is non-threatening to causal continuity: if the PCT is right, psychological constancy is simply unnecessary for the persistence of personal identity.

My first concern with this position is that it is vague with respect to what and how much has to be transmitted, with what accuracy. Suppose that at time t_1, the configuration of my brain is such that I am 99% sure that there is a jar of spaghetti sauce in the cupboard. (I assume that this would count as a belief under the PCT.) Suppose that by time t_2, the strength of the synapses encoding that belief has spontaneously decayed somewhat, lessening my confidence to 75%. Has the belief been causally transmitted enough for me to be the same person at t_2 as I was at t_1? I don't think the PCT answers that question. To answer *yes*, I am the same person at t_2 as I was at t_1, it will not be enough to say that my psychological 75% certainty was causally inherited from my earlier 99% certainty: such inheritance is guaranteed by the causal closure of the physical. To push the example to the extreme, suppose the synapses have deteriorated completely so that I have zero confidence in the spaghetti sauce. My prior belief is now altogether gone. It would certainly appear that the belief has not been transmitted from t_1 to t_2, and yet there is a clear chain of causal inheritance, at the neural level, of my entire psychology over that time span.

Indeed—and this is my second concern—I think that the PCT implies that we are immortal. In normal circumstances (barring tele-transportation, fission, etc.) a living human being is causally continuous with their future biologically and psychologically dead self. If non-branching causal continuity is *sufficient* for the persistence of identity, then I will persist indefinitely, for there is typically no

interruption to the causal chains pertaining to our animal selves, from birth to death and beyond, into total dissolution of our bodies and brains. If causal continuity is *necessary* but not sufficient, then it should be a non-trivial necessity…but the causal closure of the physical (plus psychological supervenience on the physical, which I am assuming) means that non-branching causal continuity always obtains, i.e. is trivial.

On my constancy account, the spaghetti sauce question is answered as follows: If the percent certainty of the belief is judged to be a defining property of my identity, then I do not persist over the span. If a coarser measurement of that belief, like 'Do I think it more likely than not that there is spaghetti sauce in my cupboard?' is a defining property, then I do persist. The burden of determining whether I persist falls to the person who wants to know. There is no fact of the matter. In this particular case, we suspect that beliefs concerning spaghetti sauce would be judged by almost everyone as merely incidental features of my psychology. Other more substantial beliefs, on the other hand, especially when bundled together into broad functional relations (e.g. Do I regard supernatural explanations favorably?), could be more commonly regarded as constituting definitional and could tip the balance so that if enough of them changed, I would become a different person. The former me would cease to exist because there would be no entity that computed the function that the former me computed; there would be no more tokens of the type defined by the former beliefs.

A refinement of the PCT is Nozick's theory of the closest continuer, which says the following (his ideas, my notation): Let $f(x,y)$ be a metric of the functional closeness of two entities. For all x, $f(x,x)=0$. If there is a y at t_2 such that $f(x \text{ at } t_1, y \text{ at } t_2)=0$, then x and y are identical. If there is no such y, then that entity z that minimizes $f(x \text{ at } t_1, z \text{ at } t_2)$ is identical with x at t_1. In the case of a tie, i.e. if there are two entities, z_1 and z_2, such that $f(x \text{ at } t_1, z_1 \text{ at } t_2)=f(x \text{ at } t_1, z_2 \text{ at } t_2)$ then *neither* z_1 nor z_2 is identical with x.

As is clear from the earlier quote, Nozick does not provide details about f; indeed, he does not think there is an objective standard for it, in the context of personal identity. I think he is right about that. But I think he goes wrong in other ways. First, I do not think it is necessary that there be any z at t_2 identical with x at t_1. He gives the example of the fission of the Vienna Circle into two groups, one in Istanbul, with some of the original members, and another in America with other members. Both splinter groups discuss the same things as the original group, with the same traditions and practices. Nozick says, 'Either the group in Istanbul is the Vienna Circle or isn't…' (2003, p. 95). This clearly assumes a dichotomy; at best it is a bare assumption, at worst a false one. Moreover, it is inconsistent with the theory's treatment of metric ties: if the Istanbul group and the American group were equal in closeness to the original group, it would *not* be the case that the Istanbul group either was or was not the Vienna Circle. That truth claim would be indeterminate. On my view, whether there is exactly one z at t_2 identical with x at

t_1, or no such z, or more than one, depends on what z's there are and how we define *f*.

My second problem with the closest continuer theory is that it does not do what Nozick says it does: account for our judgments of identity. At least, there are readily available cases where it does not do so. In the 'I am not the man I once was' idiom, the person I am today is, in most ordinary cases, the closest continuer of my former self. (Perhaps there are exceptions, like identical twins who share memories, beliefs, and values, one of whom changes dramatically while the other does not—but let's set these aside and focus on the canonical cases.) If my judgment of personal identity were governed by the closest continuer schema, I should not be able to mean that I am not the man I once was. But I *do* mean it. In a more exceptional case, a writer like Robert Pirsig (author of *Zen and the Art of Motorcycle Maintenance*) expresses that meaning by referring to his former self in the third person. There, an acute psychotic episode resulted in a psycho-functional discontinuity across which personal identity did not *in his judgment* persist. This cannot be, according to the closest continuer theory. 'But surely, it is still the same man,' one objects. '—the same man, deep down.' This just signals a different choice of functional constitution. On my account, there is no one right choice. When the Count of Monte Cristo says that Edmond Dantès is dead, he is telling a truth—and so is Mercédès, when she says that he still lives.

Expressions like 'Of course I couldn't answer, I was unconscious,' and 'I was asleep when you called' suggest that we think that even when our psycho-functional capacities are suspended, we still exist. A continuity theorist might explain this as follows. Clearly, sleep and unconsciousness temporarily disable the exhibition of many psychological functions, but in virtue of the brain's largely constant anatomical configuration, those functions are causally transmitted to the future self. Moreover (the continuity theorist may say) sleep and unconsciousness cause trouble for functional constancy theories, because the function computed during sleep is not the same as the function computed while conscious. This is an interesting objection. My response is as follows.

While asleep, the parts of my brain functionally relevant to many forms of psychological activity, including perhaps those the complainant above regards as constituting personal identity, cannot receive input, properly process it, or output behaviors in the usual way. In virtue of its largely constant anatomical configuration, however, my brain retains its waking functional dispositions while I sleep. Are these *counterfactual* dispositions? —or are there traces of them which a sensitive investigation could reveal? I think the latter. Suppose, for example, that the parts of my brain in question were actually missing, as opposed to merely functionally silenced by sleep processes. Surely this physical fact would have *some* consequences for the rest of my brain. It is a problem in practice, but not in principle, to expose that this is the case. In other words, we might say that while I sleep, my brain actually does compute the same function as it does while I am

awake; it's just that the normal methods of exposing that function, including of course casual observation, are too crude to discover this.

Consider, finally, the expression, 'Sorry, I'm not myself tonight.' This seems to reflect a self-model containing two superimposed personalities. The first is exhibited by the way I am acting (the function I am computing) tonight. This is asserted to be a different person than the second, which is the way I normally am. That second, typical self stands in the same relation to my typical self as does the audience of this utterance. What all of this suggests to me is that part of our normal conception of 'I' includes the model of our personality as it is seen by others—pretty much, incidentally, Sartre's *être-pour-autrui*. It's not just that I know declaratively that I am perceived by others as an agent with particular properties; it's that that third-person view of myself is baked into my self-model as part of a mixed first- and third-person chimera. I don't know how a continuity theorist would account for such expressions, but I'd like to account for it by appealing to the extended self. If our models of ourselves can include cars and Monopoly pieces, surely they can also contain third-person views of ourselves. I suspect, though I won't develop this claim here, that the neurocognitive origins of that model may be understood as emerging from the brain's dual use of body-centric and world-centric representations (Snyder et al., 1998) and other neural circuits, like mirror neurons, that support contextual representations of the self and others (Uddin et al., 2007).

In short, I think that any broad theory of personal identity has to explain the ordinary situations under which selves cease to be, and theories that make continuity rather than constancy their centerpiece are weak in this regard.

5. Conclusion

It has been my assumption that linguistic expressions mirror, to a first approximation, our concepts. For the most part, only philosophers have concepts of personal identity based on exotic circumstances like fission and brain-state transfers. It would not be especially surprising, then, to find that philosophers' concepts of personal identity differed from those of the ordinary person—and I think they do differ. At the same time, even the more common view has some interesting and perhaps unexpected properties, as evidenced by the way people use personal indexicals. I have argued that one of these properties is that normal people think of their personal identities as being extended, in the sense of the extended self thesis. I have also argued that people think that what makes them *them* is their realization of a certain set of abstract functional relationships.

In the famous identity puzzle, the ship of Theseus has been retired from active service and put on display. Part one: As planks degrade they are replaced with careful replicas. Eventually, none of the original wood remains—but the ship has been uninterruptedly recognizable. We might say that it has simply been well-maintained. Part two: Surprise! The original timbers have been re-assembled off-

site into what is now a crumbling wreck. The puzzle: which, if either, is Theseus's ship?

The theory of identity that I have been advocating in this paper says that in one sense, the identity of the ship—what makes it *Theseus's* ship—is a set of functional relationships. What is in that set? The functions that make it look a certain way? Smell a certain way? Have a certain history? There are no objectively right answers; people may differ regarding which properties and which levels of description matter. In a second sense, a thing is Theseus's ship just in case the thing realizes that function. The first sense is an assertion about a type, and the second is an assertion about a token. Demonstrative expressions about rigidly-designated entities (e.g. statements like, 'That is the ship of Theseus' and 'That is Abraham Lincoln') have both senses. When it comes to selves, neither sense is exhaustively correct. Most of the time, i.e. outside the problem cases invented by philosophers, rigidly-designated entities have only one token, and the functional idiosyncrasies of that *sui generis* token define the type. If it were not so—if there really were malfunctioning tele-transporters and brain-state duplicators and so on—we would probably have quite a different set of concepts and terms and would not consider such problem cases as puzzling as we do now.

REFERENCES

Addis, D. R., and Tippett, L. (2004), "Memory of Myself: Autobiographical Memory and Identity in Alzheimer's Disease," *Memory* 12(1): 56–74.

Clark, A., and Chalmers, D. J. (1998), "The extended mind," *Analysis* 58(1): 7–19.

Dennett, D. (1992), "The Self as a Center of Narrative Gravity," in F. Kessel, P. Cole and D. Johnson (eds.), *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum.

Kaplan, D. (1989), "Demonstratives," in J. Almog, J. Perry and H. Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press, 481–563.

Krasner, D. A. (2006), "Smith on Indexicals," *Synthese* 153(1): 49–67.

Martin, R., and Barresi, J. (2003), *Personal Identity*. Malden, MA: Blackwell Publishing Ltd.

Mount, A. (2008), "The Impurity of 'Pure' Indexicals," *Philosophical Studies* 138: 193–209.

Nozick, R. (2003), "Personal Identity through Time," in Martin, R. and Barresi, J. (eds.), *Personal Identity*. Malden, MA: Blackwell.

Parfit, D. (1995), "The Unimportance of Identity," in H. Harris (ed.), *Identity: Essays Based on Herbert Spencer Lectures Given in the University of Oxford*. New York: Clarendon Press.

Parfit, D. (1984), *Reasons and Persons*. Oxford: Oxford University Press.

Shoemaker, S. (2004), "Functionalism and Personal Identity—A Reply," *Noûs* 38(3): 525–533.

Shoemaker, S. (1999), "Self, Body, and Coincidence," *Proceedings of the Aristotelian Society* 73: 287–306.

Shoemaker, S. (1984), "Personal Identity: A Materialist's Account," in S. Shoemaker and R. Swinburne (eds.), *Personal Identity*. Oxford: Blackwell.

Snyder, L. H., Grieve, K. L., Brotchie, P., and Anderson, R. A. (1998), "Separate Body- and World-referenced Representations of Visual Space in Parietal Cortex," *Nature* 394: 887–891.

Tappen, R. M., Williams, C., Fishman, S., and Touhy, T. (1999), "Persistence of Self in Advanced Alzheimer's Disease," *Image—The Journal of Nursing Scholarship* 31(2): 121–125.

Uddin, L. Q., Iacoboni, M., Lange, C., and Keenan, J. P. (2007), "The Self and Social Cognition: The Role of Cortical Midline Structures and Mirror Neurons," *Trends in Cognitive Sciences* 111(4): 153–157.