Does cognitive psychology imply pluralism about the self?

Christopher Register

forthcoming in the *Review of Philosophy and Psychology*

Psychologists and philosophers have recently argued that our concepts of 'person' or 'self' are plural. Some have argued that we should also adopt a corresponding pluralism about the metaphysics of the self. The aim of this paper is twofold. First, I sketch and motivate an approach to personal identity that supports the inference from facts about how we think about the self to facts about the nature of the self. On the proposed view, the self-concept partly determines the nature of the self. This approach provides new justification for the recent empirical turn in the philosophy of personal identity. Second, I argue that closer examination reveals that the empirical evidence does *not* in fact support pluralism about the self. Instead, the evidence points toward a model of the self-concept as a complex web of attitudes that is disposed toward integration and unity. I ultimately suggest that this unifying disposition of the self-concept helps ground the existence of a singular self.

I. Introduction

A recurring theme in the philosophy of personal identity is that no single theory adequately captures how we talk and think about being the "same person" over time. The difficulty is that sometimes we seem to talk one way, but other times we talk another way. For example, sometimes we treat someone as being the same person even if their mind is gone, such as when we encounter a loved one in a persistent vegetative state. Other times, we might say that someone is a "different person" merely if their personality changes significantly, like in the famous case of Phineas Gage.[1]

The two prominent theories of personal identity are divided on these cases. The *neo-Lockean* theory holds that 'person' is a psychological notion, and that being the same person over time requires *psychological* continuity.[2] In contrast,

[1] In 1848, Phineas Gage suffered a traumatic brain injury that reportedly altered his personality in significant ways. The case is often discussed in both psychology and philosophy. See, e.g., O'Driscoll and Leach (1998), Strohminger and Nichols (2015), and Tobia (2015).

[2] This approach owes to Locke's characterization of a person as a being capable of self-reflective thought and of personhood as the grounds of "forensic" relations such as responsibility.

animalism holds that human persons just are human organisms, and that being the same person over time requires *biological* continuity. The neo-Lockean view has trouble allowing that someone could exist in a persistent vegetative state, but it can capture what we mean when we say that Phineas Gage is not the same person he once was.[3] Conversely, animalism fails to capture the importance of Gage's transformation, though it can explain the loved one's presence despite their lack of psychological continuity.

The ambivalence in our attitudes is remarkably robust. It turns up again and again in the philosophical literature on personal identity, and psychologists have reproduced the same or very similar conflicts in experimental settings. The data suggests that there is a robust dichotomy in how we think about people. Sometimes we think about persons as human organisms, yet at other times we treat certain psychological traits as essential.

The robustness of this ambivalence has led philosophers and psychologists alike to endorse pluralism, not only about how we *think* about personal identity, but also about *what exists*.[4] That is, many theorists conclude that talk of being the same person over time runs together multiple distinct things, including biological continuity, psychological continuity, and perhaps other things too. On the pluralist view, personal identity consists in multiple distinct conditions that usually overlap. There is a biological thing *and* a psychological thing, and they can come apart. Further, these things play distinct roles: the psychological thing grounds relations of responsibility and compensation, while the biological thing grounds relations of survival and anticipation.

However, pluralism does not sit easy with our ordinary beliefs and practices or with first-personal phenomenology. We take ourselves to be *unified* selves, living unified lives and experiencing the world from a singular perspective. So, there is reason to be dissatisfied with pluralism about personal identity. Indeed, the apparent need for pluralism is sometimes treated as a problem.[5]

In this spirit, some philosophers take the apparent unity of first-personal phenomenology, attitudes, and practices not just as *evidence* about the unity of personal identity, but in fact as the *grounds* of personal identity itself. Theorists

---

[3] Neo-Lockean views are compatible with Gage being the same person before and after transformation. What these views capture is that Gage transformed in ways that are relevant to survival and identity. So, if Gage's transformation had only been more extreme, the neo-Lockean would say he did not survive it.

[4] Sider (2001), Shoemaker (2007) and (2016), Tierney et al. (2014), Tierney (2020).

[5] Schechtman (2014), pp. 80-88.

have variously sought to ground personal identity in terms of agential ownership[6], person-directed attitudes and dispositions[7], narrativity[8], conventions[9], the unity of a 'person life'[10], and our 'I'-beliefs[11]. A common idea is that the person is *essentially* the focal point of special attitudes or practical roles. Let's call these views *attitude-dependent* theories of personal identity.

Because we feel and believe that we are unified, singular entities, attitude-dependent theories may help avoid pluralism. However, it's not clear how that would work. After all, the evidence shows that our attitudes and the corresponding practices give conflicting verdicts in many cases. So, pluralism is not automatically avoided on these views, and it's an open question whether attitude-dependent views can instead ground monism about persons.

The direct aim of this paper is to show how an attitude-dependent approach to personal identity could ground monism despite the apparent pluralism encoded in our attitudes. There are two further indirect aims: to motivate and support both the attitude-dependent approach and the recent 'empirical turn' in personal identity. I accomplish these aims by examining numerous recent empirical studies that investigate how we think about persons. I ultimately propose that monism is viable on *structured* attitude-dependent views, wherein the divergent conditions are interpreted as *parts* of the person. Our attitudes about persons are structured and functionally integrated by the self-concept. Which conditions count as parts of the self, and how those parts are organized into a whole, is partly determined by the self-concept.

The paper proceeds as follows. In the next section, I sketch and motivate the attitude-dependent approach. In section III, I review evidence from both philosophy and cognitive psychology that demonstrates the deep conflicts in our attitudes about persons. A key hypothesis is that the conflicts in our attitudes exhibit a robust *dichotomy* wherein distinct conditions are taken to play divergent practical roles. In section IV, I show how pluralism provides a cogent resolution to these conflicts. In section V, I argue that other results speak against the dichotomy, motivating a revised interpretation of the data. In section VI, I develop a model of the self-concept that explains the conflict in our attitudes while preserving both singularity and integration.

---

[6] Korsgaard (1989).
[7] Johnston (1989) and (2010), Ch. 4. See also Zimmerman (2013).
[8] Schechtman (1996).
[9] Braddon-Mitchell and Miller (2004).
[10] Schechtman (2014).
[11] Kovacs (2016) and (2020).

II. Thinking about the Self

I'll begin by sketching and motivating the attitude-dependent approach to personal identity. I will highlight advantages of the general approach over the two prominent alternatives, neo-Lockeanism and animalism.

Generally, attitude-dependent views hold that what it takes to be the same person over time depends on how the individual or the community thinks about and treats persons. Very roughly, these views hold that the conditions that are *taken* to be the identity conditions of persons *thereby are* the identity conditions of persons. With respect to identity over time, such a condition might be stated as follows:

> Attitude-Dependent Identity Over Time:
>
> Person A, considered at time t_1, is the same person as entity B, considered at t_2, only if A is regarded as the same person as B.

So, if you think and act as though psychological continuity secures your continued existence, then a future entity that is psychologically continuous with you is *in fact* you. If, on the other hand, you regard the continuation of your functioning body as securing your survival despite the loss of your distinctive psychological traits, then you could in fact survive losing those traits as long as your body continues to exist. Depending on how your attitudes are organized, different relations could realize the role of personal identity.

Note that the condition stated above is merely a necessary condition for being the same person over time. Typically, attitude-dependent views impose other constraints too. For example, both Johnston (1989) and Kovacs (2020) hold that there must also be at least a minimal degree of continuity over time and some self-reflective psychological traits at some time or other.[12]

There are several advantages of attitude-dependent theories. Some of these advantages are already present in the literature, while others are new.

One advantage is that the view allows for differences across individuals or communities. For example, it seems plausible that there could be a 'Star Trek' community: a community of people who think and act as though they can travel to distant worlds by means of teleportation. In contrast, it's *also* plausible that there could be a community of people who think and act as though "teleportation" results in death of the original person and the creation of a new person. The people of this community think and act as though survival requires having one's original human

[12] See Johnston (1989), p. 457, and Kovacs (2020), p. 2.

body, and therefore they believe that they cannot be teleported in the manner of Star Trek.

Some have argued that each community is right on its own terms.[13] The attitude-dependent approach gives a unified theory that respects these differences. For an individual in either community, part of what it takes for that individual to continue to exist is for there to be a future entity that meets the conditions that the individual regards as their survival conditions.

Allowing for individual differences in identity over time also enables an ecumenical theory of survival in the context of persistent vegetative states. Some people might think that a persistent vegetative state counts as death, whereas others may regard that condition as survival. Supposing that personal identity makes a moral difference, then we can justify differential moral prescriptions depending on the personal beliefs of the patient. For a patient who does not believe they survive in a vegetative state, there is no strong moral reason to keep their vegetative body on life support. For patients who believe otherwise, there is. The advantage of the attitude-dependent view is that it provides a satisfying justification for patient-relative moral claims.

The attitude-dependent view has advantages at the beginning of life, too. Neo-Lockean views have trouble explaining how you could have been younger than an infant, given that an infant has so few distinctive psychological similarities with you.[14] Animalism, meanwhile, entails that you existed as a collection of barely differentiated cells (because the human organism exists at that stage). It's natural to think that our identity conditions are somewhere in-between. Maybe you came into existence not long before birth, or at birth. We can explain that fact by appeal to some attitudes and/or practices, either yours or those of other members of your community. If you *treat* that newborn predecessor *as* you, then in fact it *is* you. In this way, the attitude-dependent approach gives cogent verdicts that are unavailable on the standard views.

The foregoing examples show advantages of attitude-dependent views involving the survival conditions of persons. Those conditions involve what it takes to continue to exist over time. Theories of personal identity should also specify what constitutes a person *at* a time. Here too, the attitude-dependent approach has advantages. For example, we can allow that one individual adopts a prosthetic limb

[13] See Johnston (1989) and (2010), especially Ch. 4.
[14] Neo-Lockean views are compatible with the fact that you were an infant *if* they do not require any sharing of particular psychological connections over time. However, those versions of the view allow counterintuitive examples of survival through drastic change. Cf. the discussion of Methuselah in Lewis (1976).

as an integral part of their body, whereas for another individual it counts merely as a tool. The relevant difference here plausibly stems from how each person psychologically relates to their prosthetic. One person has an array of attitudes directed at the prosthetic that makes it an especially important part of them, whereas the other person does not. The attitude-dependent view delivers this verdict.[15] Other views about the metaphysics of persons are either silent on or incompatible with this verdict.

We've seen how attitude-dependent views can give compelling verdicts about *identity over time* or about *constitution at a time*. We could provide analogous applications in the context of *modal properties*, i.e., the properties that say what a person could be or how they could have been. For now, I will set those other applications aside. The rest of the paper focuses on the traditional centerpiece of personal identity: existence over time. Unfortunately, it turns out that we have conflicting attitudes about what secures our survival.

III. The Problem of Conflicting Attitudes

I've given reasons to take the attitude-dependent approach seriously. In this section, I draw from empirical evidence to develop a problem for the view. This problem motivates pluralism, which I address in the subsequent sections.

The problem faced by attitude-dependent views is that our attitudes are not well-behaved. Empirical studies show that people give inconsistent judgments about how we persist. The evidence also shows that our judgments about different practical relations, like responsibility and anticipation, seem to track different conditions. These results pose a prima facie problem for the attitude-dependent approach. In particular, it's not clear how to extract consistent facts about the nature of the self from inconsistent attitudes.

The problem of conflicting attitudes goes back to classic thought experiments from Williams (1970). Williams develops two thought experiments that describe the same situation in two different ways, and he shows that we have conflicting intuitions across these cases. In the first description, we imagine that person A and person B are about to undergo an operation in which the psychological traits of A are put into the B-body and the psychological traits of B are put into the A-body. People tend to have the intuition that the operation is one of body-swapping: after the operation, person A will in fact wake up in the B-body, and vice versa. This

[15] Cf. Kovacs (2016).

intuition reflects a belief that the continuation of psychological traits secures personal persistence.

In the second description, the reader imagines that they ("you") will undergo a procedure that will replace all of their psychological traits with new traits. After the mind-replacement procedure, "you" will then be subjected to intense pain. The typical intuition here is that one expects to experience the pain *despite* the fact that one's distinctive psychology does not persist. And further, this intuition does not change even if we learn that our own psychology was 'implanted' into another body. This intuition reflects a belief that personal persistence is secured by bodily continuity *rather than* psychological continuity—directly contradicting the first intuition.[16]

The contradictory intuitions revealed by Williams have since been reproduced in psychological experiments. Blok et al. (2005) had participants read vignettes about "Jim" who undergoes a brain transplant procedure that either does or does not preserve memories. The participants were asked whether the transplant recipient was "still Jim", and they tended to judge that Jim survives brain transplant procedures *only* when memories are preserved.[17] These judgments plausibly reflect a belief that retaining some psychological traits is necessary for survival.

In a series of studies, Nichols and Bruno (2010) corroborated and expanded on these results. They first tested the psychological condition, confirming that participants tend to agree that the patient is still Jim after the transplant procedure only if his memories are preserved.[18] Nichols and Bruno then tested the second case of Williams (1970). They presented participants with a vignette about a mind-wipe procedure and asked whether the patient ("Jerry") would feel pain afterward. In a Yes/No forced choice measure, 72% agreed that "*Jerry* will feel the pain." These results confirmed that participants judge that someone will experience pain *despite* losing all memories.[19] However, in yet another study, participants were asked explicitly: "In order for some person in the future to be you, that person doesn't need to have any of your memories: Agree/Disagree." Over 80% of participants

[16] Ninan (2021) argues that the second intuition actually reflects an endorsement of the *simple view* of personal identity rather than the bodily continuity view. Given the evidence to follow, that nuance won't impact my argument.
[17] On a 0-9 scale, participants agreed that it was still Jim with a mean response of 6.6 when memories were preserved, versus 2.0 when memories were not preserved
[18] On a 0-9 scale, the mean responses were 5.45 when memories were preserved, versus 3.24 when memories were not preserved.
[19] Williams speculated that the 1st vs. 3rd-personal framing of the cases explained the discrepancy. These results refute that hypothesis.

disagreed, suggesting even more strongly that they endorse the claim that memories *are* required for personal identity.

The results suggest that there is a dichotomy in our judgments about how persons persist. Sometimes we judge as though *psychological continuity* is required for persistence, yet other times we judge as though *bodily continuity* is sufficient for securing persistence.

Using a different experimental paradigm, Tierney et al. (2014) provide further evidence of this dichotomy. Instead of eliciting category judgments about vignettes, Tierney and colleagues examine how different perceived continuity conditions affect *temporal discounting*. Temporal discounting is a ubiquitous time bias where people discount the anticipated value of temporally distant outcomes relative to temporally near outcomes. For example, people tend to judge that receiving $100 tomorrow is preferable to receiving $110 in a month (Frederick et al. 2002). The regularity of such judgments suggests that we discount the value of outcomes as a function of their temporal distance. That function is called the *discount rate*.

Bartels and Urminsky (2011) showed that perceived psychological connectedness (i.e., perceived similarity) to future selves affects the discount rate. Using vignettes that manipulate participants' sense of connectedness to a future self, they showed that high perceived self-connectedness reduces the discount rate, whereas low connectedness increased the discount rate. In other words, people exhibited less temporal discounting when they felt more psychologically connected to their future self. This work shows that judgments about the self are involved in practical reasoning about the future, suggesting that our attitudes can be probed with discounting tasks. Additionally, the practical judgments expressed in discounting tasks determine identity on some attitude-dependent views.

Tierney et al. (2014) used this paradigm to test whether the Williams' dichotomy is also reflected in practical judgments. The hypothesis is that if perceived psychological connectedness affects some discounting tasks but not others, then we have further evidence that there are at least two ways of thinking about the self: one that is sensitive to psychological traits, and one that is not. To test the hypothesis, Tierney and colleagues compared how perceived psychological connectedness impacts judgments of *how much punishment you would deserve for cheating in the past* versus its impact on judgments of *how anxious you would be about a future root canal.*

Results showed that some *but not all* self-involving practical judgments are sensitive to connectedness. In particular, beliefs about psychological connectedness do *not* impact the anticipated badness of future root canals. The pain is expected to

be just as bad no matter how psychologically connected the future self is to the current self.

In contrast, when participants judge the amount of punishment they deserve for cheating, then connectedness *does* make a difference. In particular, participants judge that a person deserves less punishment when they are less psychologically connected to their past self.[20] This dissociation provides more evidence for a dichotomy between psychological and bodily notions of the self.

The studies so far have explored attitudes from two angles: explicit judgments about whether someone persists, and practical judgments that implicate the attitudes about the self. From both angles, responses apparently reveal one and the same dichotomy between a psychological and a bodily notion of the self. Because the same dichotomy appears when probing attitudes from two angles, we have reason to believe that these effects are not merely superficial artifacts of experimental design.

The robust dichotomy in how we think about the self suggests that we have genuinely conflicting attitudes about what we are. This conflict poses a problem for the attitude-dependent approach. The problem is that, if we are conflicted about what we take ourselves to be, then it's hard to see how those attitudes could determine coherent, univocal conditions of personal identity. The attitude-dependent approach is motivated in part by the hope that these attitudes could adjudicate difficult questions of persistence, such as whether or not an individual can survive amnesia, severe dementia, teleportation, or mind-uploading. Given that we apparently have robust conflicts in our attitudes, it's hard to see how the approach could provide helpful, determinate answers in such cases. That is the problem of conflicting attitudes.

IV. Pluralism

Recently, some philosophers have defended pluralist views of the self and personal identity (Shoemaker 2007, Tierney et al. 2014, Shoemaker 2016, Tierney 2020). Some of the evidence taken to support pluralism is the very same evidence discussed in the last section, which shows that our attitudes seem to be organized around a plurality of distinct conditions. Indeed, pluralism about the self provides a solution to the problem of conflicting attitudes. That's because divergent attitudes are only contradictory on the assumption that they are about the same thing. If, instead, we allow that the divergent attitudes are in fact about distinct things, where one thing can exist without the other, then the conflict is resolved. So, there is

[20] Mott (2018) shows that perceived psychological connectedness affects judgments of legal punishment and moral criticism in similar ways.

reason to think that the attitude-dependent approach leads to pluralism. That is *not* my conclusion. However, before assessing pluralism, it will help to get clearer on what the view is and how the data supports it.

Personal identity has often been taken to be about a single practically important type of entity. Closer examination reveals that personal identity apparently consists in a combination of multiple distinct conditions. What's more, these distinct conditions can come apart, leading to what has been called the *problem of multiplicity*.[21] For example, it's plausible that the continuity of distinctive psychological traits is important for being responsible for past actions and perhaps for deserving compensation for past harms. In contrast, bodily continuity or continuity of the first-personal perspective is important for survival and the rational anticipation of experiences (such as anticipating pain). Given that you can have bodily persistence without the persistence of psychological traits, you can have one of these practical features (rational anticipation) without the other (desert).

Along these lines, Shoemaker (2007, 2016) argues that these distinct practical roles are grounded in a plurality of possibly diverging relations. Similarly, Tierney (2020) develops what's called the Subscript View, according to which there are multiple distinct overlapping entities—selves. The distinct selves are notated by subscripts, e.g., 'S_B' and 'S_P', where each distinct self has distinct persistence conditions (bodily, psychological, and perhaps others too). For the purposes of this paper, it doesn't matter whether we think of pluralism in terms of multiple *relations* or multiple *entities*, so I lump these views together.

Pluralism solves the problem of conflicting attitudes. In fact, the prevalence of conflicting attitudes provides a prima facie case for pluralism. Recall the dichotomous regularities in attitudes about selves: some of our attitudes appear to be organized around one type of condition (psychological continuity), whereas other attitudes appear to be organized around another type of condition (bodily continuity). The dichotomy provides some reason to think that the concept governing these attitudes, the self-concept, is either plural (i.e., is really multiple distinct concepts) or polysemous.[22]

---

[21] See Shoemaker (2007, 2016), Schechtman (2014), Tierney et al. (2014), and Tierney (2020). Multiplicity is a problem because it is counterintuitive and because it erodes the importance of persons and theories of personal identity.

[22] Conceptual polysemy occurs when there is a single concept with multiple senses. There is some direct evidence that the self-concept is polysemous. Knobe (2022) conducted a study in which participants tended to agree with the following "dual character" statement about a person: "There's a sense in which the man after the accident is clearly still Phineas, but ultimately, if you think about what it really means to be Phineas, you'd have to say that he is not truly Phineas at all."

Next observe that our attitudes about one practical relation (deserving punishment) are sensitive to psychological continuity, whereas attitudes about another practical relation (pain anticipations) are sensitive to bodily continuity. This pattern suggests that people think that distinct conditions have different practical significance. The natural next step is to conclude that there are distinct concepts of the self or distinct senses of the self-concept that track these distinct conditions respectively. On that view, questions about different kinds of practical relation elicit different self-concepts. Figure 1 depicts a model of this pluralist view of the psychology of the self-concept.

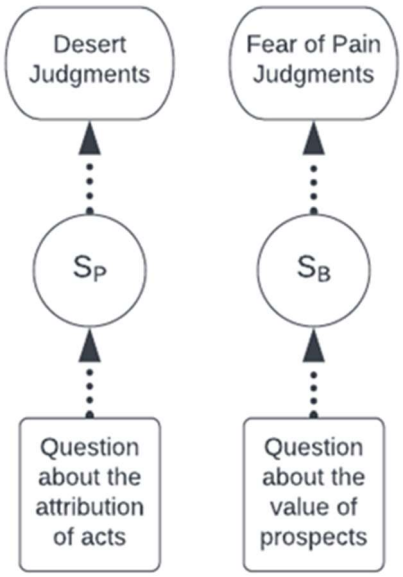

Fig. 1 *The Psychological and Bodily Self-Concepts*

On that view, the conflict in our attitudes is merely apparent, owing to the fact that we don't have distinct mental or linguistic terms by which to signal that we are thinking or talking about two different things. To render the resolution more apparent, we might (following Tierney) begin "subscripting" our language about selves. Then, there is not even an apparent conflict. Or we might continue using language in the same way while implicitly recognizing polysemy in the terms 'person', 'self', and so on.

While pluralism about the self does indeed solve the problem of conflicting attitudes, I think it misses important features of the self-concept. In particular, I believe there is a single concept that organizes and is disposed to integrate all of our various attitudes about persons. I don't deny that the self-concept is polysemous, but I do maintain that the distinct senses or bodies of information encoded in the self-concept are about just one thing. As a result, the attitude-

dependent approach can maintain monism about the self. To see why, we will have to reexamine the empirical evidence.

V. The Self-Concept is not (merely) Dichotomous

In this section, I propose that our conflicting attitudes are partial projections of a single, imperfectly unified self-concept. The fact that our attitudes diverge reveals not the presence of distinct self-concepts, but distinct aspects of one self-concept. We can then preserve the integrative structure of the self-concept, which lends itself in turn to the unity of the self. On this version of the view, distinct aspects of the self-concept are about distinct *parts* of the self.

The primary motivation for resisting pluralism comes from the observation that our intuitive sense of self is deeply monistic. Our monistic sense of self is presumably generated and reinforced, in part, by the fact that a human person tends to be a spatiotemporally continuous entity that typically has a unified field of phenomenal consciousness. But I also think there is more to this sense of unity: we are *agentially* unified in that we are motivated to satisfy all of our various drives and desires while forced to live out one coherent course of action.[23] Our internal drives and external circumstances pressure us to develop a unified conception of what we are, what we care about, and what kind of life we aim to lead. I suggest that the self-concept is disposed to encourage and maintain unity in the face of this practical pressure, thereby unifying distinct parts into an integrated whole.

We can lend support to this intuitive gloss by appealing to further empirical evidence about the behavior of the self-concept.

When we focus on the kinds of judgments involved in the Williams cases, the studies of Nichols and Bruno (2010), and that of Tierney et al. (2014), there is a risk of missing the forest for the trees. There is a risk of only seeing two apparently dichotomous patterns: a pattern of judgments oriented around the persistence of psychological traits, and a pattern of judgments oriented around bodily continuity. But those are just two patterns among many, and the fuller picture supports the idea that there is just one forest, so to speak: there is just one (complex and multifaceted) self-concept, which variously manifests in the two aforementioned ways and perhaps other ways too.

How could a fuller picture mitigate conflict? Won't there simply be *more* conflict? I agree that there will likely be more conflict, but the deeper point is that we

[23] As Korsgaard (1989) says, "In order to carry out a rational plan of life, you need to be one continuing person. You normally think you lead one continuing life because you are one person, but…the truth is the reverse. You are one continuing person because you have one life to lead."

shouldn't interpret particular judgments as free-standing. Rather, the judgments are merely *partial* projections of the self-concept.

Before turning to new studies, it's worth highlighting features of the previous studies that undermine their support for a dichotomous self-concept. First, in Blok et al. (2005), there was in fact an additional pair of test vignettes used, describing a *memory download* instead of a brain transplant. The authors do not provide data for that particular experiment. While they note that the pattern of responses is similar to the brain transplant version, they also acknowledge that "brain transplants produced higher overall ratings than memory downloads" (p. 144). Note that brain transplants preserve a higher amount of physical continuity than memory downloads. So, the discrepancy in responses suggests that while continuity of memory is important, participants are *not fully screening off* information about physical continuity. That is, the very responses that are taken to support the operations of a distinctively psychological concept of the self are *also* sensitive to bodily continuity.

Second, in both the Blok et al. (2005) and Nichols and Bruno (2010) studies, the measures were agreement scales or (in the later studies of Nichols and Bruno (2010)) forced choice questions. These measures make it hard to understand what participants have in mind when they are responding. Given the slight discrepancy in responses across the transplant and download conditions, it's plausible that the responses do not *reflect* sharp dichotomous judgments but rather that the measures *induce* dichotomous responses. In reality, the sensitivities of the self-concept are complex and graded.

Now, let us turn to new studies that further complicate the picture. Berniūnas and Dranseika (2016) showed that, in response to a vignette about someone named 'Deivydas' who gets into a car accident and enters a persistent vegetative state, participants tend to agree that "the patient is still Deivydas" and that "the patient is still a person."[24] These results show that participants sometimes judge as though being the same person does not require the preservation of psychological traits, even when they are *not* focused on anticipating future experiences.

Berniūnas and Dranseika (2016) also conducted a version of the brain transplant study, except they coached some participants on the difference between *qualitative similarity* and *numerical identity*. As others have pointed out, phrases such as "same person" or even "still Jim" are ambiguous between mere similarity and numerical identity (Starmans and Bloom 2018, Finlay and Starmans 2022). When participants

[24] The responses were above the midpoint of 4.5 in all but one sub-condition, which had a mean response of 4.38.

were asked whether the patient after the transplant was *numerically* the same person as before the procedure, the loss of memories from the procedure did not have a significant impact on their judgments. This result further undermines the claim that the earlier responses reflect beliefs about a strictly psychological notion of self or person. It could instead be that the psychological traits are believed to be an important part of the person, rather than the sole basis of their continued existence.[25]

Other studies undermine another claim of Tierney et al. (2014): that temporal discounting effects reflect the same dichotomy in our attitudes. Recall that psychological connectedness affects discounting with respect to punishment judgments, but not with respect to anticipations of pain. Tierney et al. (2014) suggest that this discrepancy reflects the same dichotomy that we saw in the Williams cases. However, studies show that a variety of factors affect discounting. And, contrary to what was suggested by the pain discounting task, perceived psychological connectedness can affect how people evaluate future experiences.

As an example, like the Bartels and Urminsky (2011) studies already discussed, Bartels and Rips (2010) show that higher perceived psychological connectedness results in a lower rate of temporal discounting. These studies also showed that participants preferred benefits to occur *before* large changes in connectedness and for costs to occur *after* such changes. That suggests that perceived psychological continuity does sometimes affect the anticipated value of future experiences, contrary to the conclusion that reasoning about prospective experiences engages a strictly bodily conception of the self.[26] So, there is no sharp dichotomy in discounting judgments, either.

Given that perceived psychological similarity affects how we reason about future experiences, the discrepancy highlighted by Tierney et al. (2014) is likely due to peculiarity in how *pain* is evaluated in practical reasoning. If that's right, then we do not need to posit distinct self-concepts operating in different types of practical reasoning. That, in turn, undermines the claim that it is a *different entity* (or relation) that we have in mind when we reason about pain as compared to reasoning about, e.g., moral responsibility or punishment.

---

[25] Other work, like Tobia (2015), shows that participants are *not* merely making similarity judgments. Tobia's results show that the *direction* of change matters, but qualitative similarity by itself only captures the *degree* of change.

[26] See Hershfield et al. (2009), Hershfield et al. (2011), Joshi and Fast (2013), and Hershfield et al. (2018) for similar effects of perceived psychological connectedness on practical judgments. Note that these studies, as well as that of Bartels and Rips (2010), use the 'Inclusion of Others in Self' scale (Aron, Aron, & Smollan 1992), which has participants select from seven pairs of increasingly overlapping circles to indicate how "connected and similar" they feel to the future self.

In this section, we have seen evidence that the self-concept is not sharply dichotomous. Contrary to the 'division of labor' depicted in Figure 1, some judgments about the value of future experiences are sensitive psychological continuity, and some judgments about attribution or about whether someone is the "same person" are sensitive bodily continuity. So, I posit an imperfectly unified self-concept: a singular complex web that is active in all self-involving judgments and yet is capable of partial projections. Figure 2 depicts the general model (left) as well as two partial projections that can be generated:

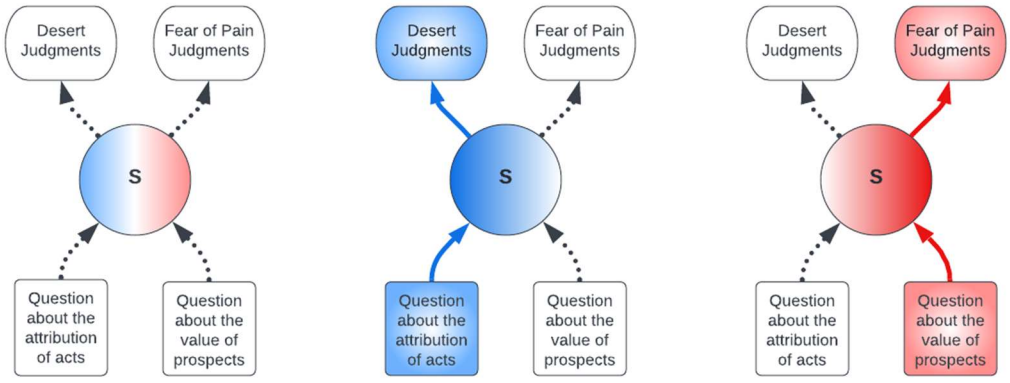

Fig. 1 *The Complex Self-Concept (left) and two activation patterns (center and right)*

In the center model, we see a possible deployment of the self-concept in response to inputs about attribution. Questions about attribution elicit a *partial activation* of the self-concept which then produces an output skewed by that partial activation: a *partial projection*. This projection corresponds to the mind-swapping case of Williams (1970) or the punishment judgments of Tierney et al. (2014) that raise the psychological traits of the self to attention.

On the righthand model, we see another possible deployment of the self-concept, meant to capture what goes on when we respond to vignettes about the anticipation of pain. Questions about pain anticipation and perhaps other forms of prospective evaluation elicit a different partial activation of the self-concept. In turn, those questions elicit projections that express the importance of bodily or bare perspectival continuity in the overall conception of the self.

The two partial activations and their projections show how an imperfectly unified self-concept could produce apparently dichotomous judgments in response to particular elicitation conditions. The self-concept is capable of partial activation

because it encodes a complex web of beliefs and bits of information, where some stimuli elicit some but not all of the encoded beliefs or bits of information.[27]

VI. The Integrated Self

I have argued that closer inspection of the evidence supports monism rather than pluralism about the self-concept. What pluralists interpret as a judgment about the *biological self*, I am suggesting should be interpreted as a judgment about the biological aspects or parts of the self. To support that interpretation, I demonstrated that a model of *partial activations* can explain the divergent attitudes produced by the self-concept. A partial activation raises just some parts of the self to salience, and downstream attitudes are then guided by the activated representation of those features. Different parts of the self-concept generate different downstream attitudes, and that's why we see conflict.

However, that does not yet explain how to extract univocal identity conditions from the conflicted behavior of the self-concept. More specifically, you may find yourself still wondering whether your loved one can survive with severe dementia or in a persistent vegetative state. How does the monistic model of the self-concept help the attitude-dependent view resolve these questions?

The problem is that your loved one will probably recapitulate the conflicting attitudes canvassed above. The solution is that their self-concept *also* provides the basis for resolving this conflict. The self-concept encodes information that is not revealed by the conflicting attitudes, and this additional information provides a non-arbitrary basis for a kind of idealization that delivers us from conflict. Importantly, the activity of the self-concept is also largely governed by a presumption of singularity, i.e., a belief that there is just a singular self. The presumption of singularity generates a disposition to resolve the conflicts rather than to bifurcate one's beliefs and concepts. With these resources in hand, we can then say that your loved one's identity conditions are the conditions they *would* take to be their identity conditions if they were able to fully integrate the beliefs encoded in their self-concept. So, we resolve conflict by appealing to the integrative disposition of the individual's self-concept.

---

[27] Atomist views according to which concepts have no internal structure are compelling (Fodor 1998, Quilty-Dunn 2020). One might worry that my view of the self-concept as a complex web is at odds with atomist views. On the contrary, my view is compatible with the claim that concepts are atomic representations that simply activate bodies of information that are not internal to the concept. On that picture, the mental item that is capable of partial activation is more like a self-*conception* rather than the self-concept on its own.

That's the solution in broad strokes. But what evidence is there that we have beliefs that can support integration? And how should we expect the resolution to shake out?

I'll briefly mention two further paradigms in the cognitive psychology of the self-concept that could support and motivate conflict resolution, but a full answer to this question requires further empirical work. In particular, cognitive psychologists should continue to investigate not just *what* we believe about ourselves, but also how those beliefs fit together and interact.

The first paradigm is that the self-concept has been shown to encode beliefs about the *relative importance* of various parts of the self. For example, several results show that we tend to report that changes in our moral traits and distinctive personality traits are *more* disruptive of our identity than changes to other traits, such as perceptual capacities (Strohminger and Nichols 2014, Heiphetz et al. 2017). It's natural to interpret these results as revealing that we take some features to be more important to what we are, or to our continued existence, than other features. This kind of belief is poised to play a role in how we revise in the face of conflicting attitudes. For example, if someone were to point out that moral traits and perceptual traits come apart in the Williams' mind-swapping case, that may sway participants to side with whatever conditions secure what matters more to them, i.e., their moral traits.

The second paradigm is that the self-concept has been shown to encode beliefs about the *causal structure* of the self (Chen et al. 2016). Interestingly, the causal structure of the self in many ways mirrors the relative importance of the features just discussed. The results show that some traits ascribed by the self-concept are believed to be more *causally central* than others, in the sense that they cause those other traits. Additionally, changes in the traits that are judged to be more causally central are also judged to be more disruptive to identity than changes in traits that are causally peripheral. So, the causal beliefs encoded in the self-concept bear on identity judgments. These causal beliefs could also drive conflict resolution in the self-concept. For example, if I believe two traits are part of what makes me who I am, yet I believe that one of these traits is more causally central than the other, then we should expect that I am disposed to jettison the belief about the causally peripheral trait rather than the belief about the causally central trait.

One specific avenue for future empirical work is to investigate *dependencies* in the self-concept. Sometimes when we have multiple beliefs about a subject matter, one of the beliefs is held *because of* another belief. For example, I believe that oranges are healthy because I believe they contain vitamin C and also believe that vitamin C is healthy. If I found proof that vitamin C was actually unhealthy, then I would

be disposed (perhaps when reflecting properly) to abandon my belief that oranges are healthy. That's because the latter belief depends on the belief that vitamin C is healthy.

Similarly, dependency among beliefs could guide conflict resolution in the self-concept. For example, suppose I believe that my survival requires the continuation of my original body *only* because I believe that my body secures the continuation of my first-personal perspective. Because that belief is dependent—because I do not assign any independent importance to being a biological organism but only assign importance to it *derivatively*—then there is a sense in which that attitude is not a core part of my self-concept. What *really* counts is the non-derivative commitment, i.e., my belief that I have a first-personal perspective. Moreover, the dependency structure determines how I would revise my beliefs in the face of apparent conflict. Though I sometimes report that my body secures my survival, in the face of conflict I would jettison the belief that I cannot survive without my body.

This internally supported standard of idealization helps resolves conflict. In general, a person will tend not to tolerate conflicts in how she thinks about herself. That intolerance—the disposition toward coherence—is itself a constitutive feature of the monistic and integrative nature of the self-concept. That disposition toward coherence is a psychological disposition to treat various conditions *as* belonging to one and the same object. And it is that disposition that pushes each of us toward a unified, univocal self-conception.

One important upshot of this picture of integration is that the determinate nature of the self ultimately depends on features to which the self-concept is sensitive, even if someone doesn't currently recognize those features. There is evidence from cognitive psychology that shows that the self-concept is sensitive to features that are external to the person themselves. For example, a host of evidence shows that evaluative and or normative elements guide our judgments about ourselves.[28] Other evidence suggests that social relations play a similar role.[29] If that's right, then the self-concept approach may end up producing selves that are partly constituted by their relations to other people, events, social facts, values, and norms. That implication suggests that the self-concept approach may help unify non-standard views in the study of the self, such as narrative[30] or relational[31] conceptions.

---

[28] For representative work across various paradigms, see Strohminger, Knobe, and Newman (2017), Tobia (2015), and Chen, Urminsky, and Bartels (2016).
[29] Heiphetz, Strohminger, and Young (2017).
[30] Schechtman (1996, 2014), Taylor (1985).
[31] Dover (2022) has recently explored ways that conversation can determine the self. See also Husserl (1960), especially the 5th meditation (pp. 89-151) for an early phenomenological analysis that also reveals dependence of the self on others. Butler

The final upshot of the foregoing view is what it reveals about the problem of multiplicity. The problem of multiplicity is the worry that persons are really just an amalgam of multiple distinct and possibly divergent relations, each of which is of differing practical importance (Shoemaker 2007, 2016).

One point that I have defended is that the self-concept is monistic because it exemplifies an overarching disposition toward unity and integration. That very same disposition, I suggest, shows what is wrong in the concession that personal identity is really about multiple distinct things. First, observe that we do not regularly think or talk about the self in pluralist terms. Part of the reason for this is that we have a deep-seated monistic sense of self—a sense that reflects our disposition toward coherence in self-involving actions and beliefs. That mental property of us, that very disposition, is a real relation that ties together our various practical roles. That is, there is a higher-order coordination across these roles that *makes* it the case that there is a unified entity at the hub of those roles. So, the unity of the self-concept helps solve, not only the conflicts in our attitudes, but also the problem of multiplicity—that is, *if* we are willing to take on board the idea that the unifying dispositions of the self-concept are a constitutive part of the self.

References

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Bartels, D. M. & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *J. Exp. Psychol. Gen.* 139(1), 49-69. https://doi.org/10.1037/a0018062

Bartels, D. M., & Urminsky, O. (2011). On intertemporal selfishness: How the perceived instability of identity underlies impatient consumption. *Journal of Consumer Research, 38*(1), 182–198. https://doi.org/10.1086/658339

Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 127–149). Washington, DC: Am. Psyc. Association.

(2005, pp. 65-101) develops similar themes, drawing from both phenomenological and psychoanalytic traditions.

Braddon-Mitchell, David & Miller, Kristie (2004). How to be a conventional person. *The Monist* 87 (4):456-474.

Butler, Judith (2005). *Giving an Account of Oneself*. New York: Fordham University Press.

Chen, S. Y., Urminsky, O., & Bartels, D. M. (2016). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science*, 27, 1398–1406.

Dover, Daniela (2022). The Conversational Self. *Mind*, Volume 131, Issue 521, pp. 193–230, https://doi.org/10.1093/mind/fzab069

Finlay, M., & Starmans, C. (2022). Not the same same: Distinguishing between similarity and identity in judgments of change. *Cognition*, *218*, 104953. https://doi.org/10.1016/j.cognition.2021.104953

Fodor, Jerry A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40, 351–401.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity. *Cognitive science*, 41(3), 744–767. https://doi.org/10.1111/cogs.12354

Hershfield, H.E. (2011), Future self-continuity: how conceptions of the future self transform intertemporal choice. Annals of the New York Academy of Sciences, 1235: 30-43. https://doi.org/10.1111/j.1749-6632.2011.06201.x

Hershfield, H.E., Garton, M. T., Ballard, K., Samanez-Larkin, G. R., & Knutson, B. (2009). Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving. *Judgment and dec. making*, *4*(4), 280–286.

Hershfield, H.E., Goldstein, D.G., Sharpe, W.F., Fox, J., Yeykelvis, L., Carstensen, L.L., & Bailenson, J. (2011). Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research*, 48, S23-S27.

Hershfield, H., John, E. M., Reiff, J. S. (2018). Using Vividness Interventions to Improve Financial Decision Making. *Policy Insights from the Behavioral and Brain Sciences*, 5(2): 209-215. DOI: 10.1177/2372732218787536

Husserl, Edmund (1960). *Cartesian Meditations: An Introduction to Phenomenology*. Translated by Dorion Cairns. Kluwer Academic.

Johnston, Mark (1989). Relativism and the Self. In M. Krausz (ed.), *Relativism: Interpretation and Confrontation*. Notre Dame University Press.

Johnston, Mark (2010). *Surviving Death*. Princeton University Press. Princeton, NJ.

Joshi, P., & Fast, N. (2013). Power and reduced temporal discounting. *Psychological science*, *24*(4), 432–438. https://doi.org/10.1177/0956797612457950

Knobe, Joshua (forthcoming). Personal Identity and Dual Character Concepts. In K. Tobia (ed.), *Experimental Philosophy of Identity and the Self*. Bloomsbury.

Korsgaard, Christine M. (1989). Personal identity and the unity of agency: A Kantian response to Parfit. *Philosophy and Public Affairs* 18 (2):103-31.

Kovacs, David (2016). Self-made People. *Mind* 125 (500):1071-1099.

Kovacs, David (2020). Diachronic Self-Making. *Australasian J. of Philosophy*:1-14.

Lewis, David K. (1976). Survival and identity. In Amelie Oksenberg Rorty (ed.), *The Identities of Persons*. University of California Press. pp. 17-40.

Longenecker, Michael (2021). Community-Made Selves. *Australasian J. of Philosophy*.

Molouki, S., Chen, S.Y., Urminsky, O. & Bartels, D.M. (2020). How personal theories of the self shape beliefs about identity continuity. In Lambert, E. & J. Schwenkler (Eds.) *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*. Oxford: Oxford University Press.

Mott, Christian (2018). Statutes of Limitations and Personal Identity. In Tania Lombrozo, Joshua Knobe & Shaun Nichols (eds.), *Oxford Studies in Experimental Philosophy, Volume Two*. New York, NY, USA: pp. 243-269.

Nichols, S. & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, 23:3, 293-312, DOI: 10.1080/09515089.2010.490939

Ninan, Dilip (2021). Williams on the self and the future. *Analytic Philosophy*.

O'Driscoll, K., & Leach, J. P. (1998). "No longer Gage": an iron bar through the head. Early observations of personality change after injury to the prefrontal cortex. *BMJ (Clinical research ed.)*, 317(7174), 1673–1674. https://doi.org/10.1136/bmj.317.7174.1673a

Quilty-Dunn, Jake (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind and Language* 36 (1):158-185.

Berniūnas, Renatas & Dranseika, Vilius (2016) Folk concepts of person and identity: A response to Nichols and Bruno, *Philosophical Psychology*, 29:1, 96-122, DOI: 10.1080/09515089.2014.986325

Schechtman, Marya (1996). *The Constitution of Selves*. Cornell University Press.

Schechtman, Marya (2014). *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford University Press.

Shoemaker, David (2007). Personal Identity and Practical Concerns. *Mind*, 116 (462): 317-357.

Shoemaker, David (2016). The Stony Metaphysical Heart of Animalism. In Stephan Blatti & Paul Snowdon (eds.), *Animalism*. Oxford Uni. Press. pp. 303-328.

Sider, Theodore (2001). Criteria of Personal Identity and the Limits of Conceptual Analysis. *Philosophical Perspectives* 15:189-209.

Starmans, C., & Bloom, P. (2018). Nothing Personal: What Psychologists Get Wrong about Identity. *Trends in cognitive sciences*, *22*(7), 566–568. https://doi.org/10.1016/j.tics.2018.04.002

Strohminger, N., Knobe, J., & Newman, G. (2017). The True Self: A Psychological Concept Distinct from the Self. *Perspectives on Psychological Science*, 12(4), 551–560. https://doi.org/10.1177/1745691616689495

Strohminger, Nina & Nichols, Shaun (2014). The Essential Moral Self. *Cognition* 131 (1):159-171.

Strohminger, Nina & Nichols, Shaun (2015). Neurodegeneration and identity. *Psychological Science* 26 (9):1469– 1479.

Taylor, Charles. (1985). Self-interpreting animals. In *Philosophical Papers*, pp. 45-76. Cambridge University Press. doi:10.1017/CBO9781139173483.003

Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. (2014). How Many of Us Are There? In Justin Sytsma (ed.), *Advances in Experimental Philosophy of Mind*. Continuum Press.

Tierney, Hannah (2020). The Subscript View: A Distinct View of Distinct Selves. In *Oxford Studies in Experimental Philosophy*. pp. 126-323.

Tobia, Kevin P. (2015). Personal identity and the Phineas Gage effect. Analysis 75 (3):396-405

Williams, B. A. O. (1970). The self and the future. *The Philosophical Review*, 79(2), 161–80.

Zimmerman, Dean (2013). Personal identity and the survival of death. In F. Feldman and B. Bradley (ed.), *The Oxford Handbook of Philosophy of Death*. pp. 97-154.