Me, My (Moral) Self, and I

Jim A.C. Everett, Joshua August Skorburg, Jordan L. Livingston

In this chapter we critically review interdisciplinary work from philosophy, psychology, and neuroscience to shed light on perceptions of personal identity and selfhood. We review recent research that has addressed traditional philosophical questions about personal identity using empirical methods, focusing on the "moral self effect": the finding that morality, more so than memory, is perceived to be at the core of personal identity. We raise and respond to a number of key questions and criticisms about this work. We begin by considering the operationalization of identity concepts in the empirical literature, before turning to explore the boundary conditions of "moral self effect" and how generalizable it is, and then reflecting on how this work might be connected more deeply with other neuroscience research shedding light on the self. Throughout, we highlight connections between classical themes in philosophy, psychology, and neuroscience, while also suggesting new directions for interdisciplinary collaboration.

Everett, J.A.C, Skorburg, J.A., & Livingston, J. (Forthcoming). Me, my (moral) self, and I. In Sinnott- Armstrong & De Brigard (Eds) *The Handbook of Philosophy and Neuroscience*. Cambridge, MA: MIT Press.

Me, My (Moral) Self, and I

What makes you the same person you were five years ago? Will you be the same person that will exist in five years time? Would you be the same person if, by some fantastical freak of nature, you woke up in someone else's body? These questions have fascinated humans for thousands of years, and with the specter of new technological advances such as human-machine cyborgs looming, they will continue to do so. They will continue to do so because they get at a fundamental human question: what makes *you*, you?

In this chapter we outline the interdisciplinary contributions that philosophy, psychology, and neuroscience have provided in the understanding of the self and identity, focusing on one specific line of burgeoning research: the importance of morality to perceptions of self and identity.[1] Of course, this rather limited focus will exclude much of what psychologists and neuroscientists take to be important to the study of self and identity (that plethora of self-hyphenated terms seen in psychology and neuroscience: self-regulation, self-esteem, self-knowledge, self-concept, self-perception, and more (Katzko, 2003; Klein, 2012). We will likewise not engage with many canonical philosophical treatments of self and identity. But we will lay out a body of research that brings together classic themes in philosophy, psychology, and neuroscience to raise empirically tractable philosophical questions, and philosophically rigorous empirical questions about self and identity.

More specifically, in Section 1, we will review some recent research which has treated traditional philosophical questions about self and identity as empirical questions. Within this body of work, we will be primarily concerned with the finding that morality (more so than memory) is perceived to be at the core of self and identity. Then, in Section 2, we raise and respond to a variety of questions and criticisms: first, about the operationalization of identity concepts in the empirical literature; second, about the generalizability of the "moral self effect"; third about the direction of change; fourth, about connections with recent work in neuroscience; fifth, about the target of evaluation. Finally, in Section 3, we consider a variety of implications and applications of this work on the moral self. Throughout, we aim to highlight connections between classical themes in philosophy, psychology, and neuroscience, while also suggesting new directions for interdisciplinary collaboration.

[1] In this chapter, we use "self and identity" as a catch-all for these various approaches. However, when quoting other research or engaging with specific criticisms, terms such as "personal identity", "identity persistence", etc. will inevitably crop up as well. But unless otherwise noted, we take ourselves to be primarily engaging with the (relatively narrow) literature, described in more detail below, which examines folk judgments of identity change.

1 From the Armchair to the Lab

Given our limited focus, we follow Shoemaker (2019) in dividing contemporary philosophical approaches to morality and personal identity into four broad categories: Psychological views, biological views, narrative views, and anthropological views. For present purposes, much of the research we describe below is situated within the domain of psychological views. This family of views about personal identity generally holds that person *X* at Time 1 is the same person as *Y* at Time 2 if and only if *X* is in some sense psychologically continuous with Y (henceforth, "psychological continuity" views). There is much debate, of course, about what such psychological continuity consists of, but following Locke (1975), memory is traditionally thought to be central in this regard. So, *X* at Time 1 is the same person as *Y* at Time 2 if *X* can, at one time, remember an experience *Y* had at another time. This vague formulation is subject to a number of objections (see Olson 2019 for an overview), but our concern here is not to vindicate or impugn any particular version of a psychological theory of personal identity. Instead, we want to focus on the ways in which such philosophical theorizing might be informed by work in psychology and neuroscience - and also how work in psychology and neuroscience can shape philosophical theorizing.

As Tobia & Shoemaker (forthcoming) point out, if a psychological continuity view depends on a certain view of memory or consciousness, then experimental methods might eventually be able to clarify the precise mechanisms which support the relevant kinds of memory or consciousness. But by and large, this is not the preferred strategy in the literature. Instead, much of the relevant interdisciplinary work has focused on *judgments of identity change* and the psychological processes that underlie such judgments, and our work is a contribution to this trend.

In this vein, some of the early work in experimental philosophy took psychological continuity views as testable hypotheses. For example, Blok, Newman, & Rips (2005) asked participants to imagine that a team of doctors removed a patient's brain (call him "Jim") and destroyed his body. Then, in one version of the vignette, participants were told that the doctors successfully transplanted Jim's brain into a new body, such that the memories in Jim's brain are preserved in the new body. In another version of the vignette, participants were told that the memories were not preserved during the transplant.

In these kinds of scenarios, psychological continuity theories predict that in cases where *memories* were preserved, participants should judge that the brain recipient is "still Jim," where in cases where the memories are not preserved, participants should disagree that the brain recipient is "still Jim." And indeed, this is precisely what Blok, Newman, & Rips (2005) found, and it was also replicated by Nichols & Bruno (2010). There are, as we will see below, concerns about the adequacy of this methodology. Still, these studies helped to spark a line of research in which armchair philosophical speculation about self and identity formed the basis of empirical inquiry. In turn, this methodology opened a range of new questions about specific components of psychological continuity views. Is

there something special or unique about memory? Or might other psychological features be important for understanding judgments about self and identity?

1.1 The unbearably moral nature of the self

Even in the earliest accounts of psychological continuity, morality has played an important role. After all, for Locke, personal identity could also be understood as "forensic" concept: it is tied to memories, but it is also crucial for moral concepts such as responsibility, praise and blame.

In recent years an influential line of work has suggested that *morals*, more so than memories, are actually perceived to be at the heart of self and identity. Sparking this line of research, Prinz and Nichols (2017), Strohminger and Nichols (2014) and others, presented participants with a wide variety of traits and asked them to imagine how much a change to a specific trait would influence whether someone is perceived to be the same person (e.g. "Jim can no longer remember anything that happened before the accident. Aside from this, he thinks and acts the same way as before the accident. Is Jim still the same person as before the accident?").

To test the importance of different kinds of psychological traits, Strohminger and Nichols (2014) used a variety of items that detailed different changes that a person could go through. Some were moral changes (e.g. now Jim is a psychopath or pedophile); others were personality changes (e.g. now Jim is shy, or absentminded); others were a loss of memories (e.g. now Jim cannot remember time spent with parents, or cannot remember how to ride a bike); others were changes to desires and preferences (e.g. now Jim desires to eat healthily, or wants to quit smoking); and yet others were perceptual (e.g. now Jim has lost his ability to feel pain or see color).

According to classic psychological continuity views that prioritize memories, one should expect changes to memories to be more impactful on perceived identity persistence than other kinds of changes. Instead, across the various studies and specific items, results have consistently supported a "moral self effect": when someone changes in terms of moral traits like honesty, empathy, or virtuousness, they are rated as more of a different person than when they change in terms of memories, preferences, or desires. Thus, "moral traits are considered more important to personal identity than any other part of the mind" (Strohminger & Nichols 2014, p. 168).

This finding is not a one-off: The moral self effect holds across a wide variety of scenarios (Strohminger, Knobe & Newman, 2016; Prinz & Nichols, 2017) and has been replicated across contexts. For example, the effect has been replicated across age groups, such that 8 to 10- year-olds rate moral changes as most core to identity (Heiphetz, Strohminger, Gelman, & Young, 2018), as well as across cultures, such that Buddhist monks in India also rate moral changes as most core to identity (Garfield, Nichols, Ray, & Strohminger, 2015). The moral self effect has even been replicated in real-world contexts such that family members of patients with neurodegenerative diseases tend to rate changes to moral faculties as more disruptive to identity than changes to memories

(as in Alzheimer's disease) or changes to physical motor functions (as in amyotrophic lateral sclerosis) (Strohminger & Nichols, 2015). In fact, we know of only one group of individuals in which the moral self effect does *not* appear to replicate — in psychopaths (Strohminger, 2018).

2 Some Questions and Criticisms from Philosophy, Psychology, and Neuroscience

Having reviewed some of the relevant work on the moral self effect, we now turn to in-depth considerations of some questions and criticisms of this research program. First, we will consider prominent criticisms about the operationalization of identity concepts in the moral self literature and we will detail some responses based on our recent work. Next, we raise questions about the applicability and generalizability of the moral self effect. Then, we will consider questions about the direction of change and connections with related research on the "true self". Finally, we draw novel connections with recent work in neuroscience. Throughout, given the interdisciplinary nature of the topic, we argue that philosophy, psychology, and neuroscience can each contribute to addressing these questions, albeit to varying degrees. In the section that follows, we review the ways in which each discipline has contributed to addressing these questions, noting the strength and weaknesses of each approach, and incorporating a review of our own interdisciplinary work along the way.

2.1 Is Identity Quantitative, Qualitative, or Both?

A prominent question that has been raised in response to the findings canvassed above is whether, when faced with these thought experiments, participants are really reporting that someone has become a *different person*, or if instead, they are just reporting *dissimilarity*. (Berniunas & Dranseika, 2016; Dranseika, 2017; Starmans & Bloom, 2018). In the philosophical literature (something like) this idea has been variously conceptualized as the difference between quantitative identity on the one hand, and qualitative identity, on the other (Parfit, 1984; Schechtman, 1996).

When faced with a question like "Is *X* the same as before?", Schechteman (1996) notes that there are two different approaches. The first concerns re-identification: What are the necessary and sufficient conditions under which *X* at Time 1 is identical to *Y* at Time 2? These kinds of questions are about the logical relation of identity, and are often discussed in terms of quantitative or numerical identity. The second concerns characterization: What makes *X* the person that they are? These questions are about which actions, beliefs, values, desires, and traits are properly attributable to a person, and are often discussed under the heading of qualitative identity. Similarly, Parfit (1984) highlights the importance of distinguishing qualitative vs. numerical identity:

> There are two kinds of sameness, or identity. I and my Replica are qualitatively identical, or exactly alike. But we may not be numerically identical, or one and the same person. Similarly, two white billiard balls are not numerically but may be qualitatively identical. If I paint one of these balls red, it will cease to be qualitatively identical with itself as it was. But the red ball that I later see and the white ball that I painted red are [numerically] identical. They are one and the same ball. (p 201)

When it comes to persons, twins, for example, can be qualitatively identical, but still numerically distinct people: you might not be able to tell them apart, but they still need two passports to travel abroad. Variants of this distinction have been used to criticize the early studies on the importance of memories to identity persistence (e.g. Blok, Newman, & Rips, 2005, 2007; Nichols and Bruno, 2010).

As Berniunas and Dranseika (2016) argue, these experimental designs potentially conflate qualitative and numerical concepts of identity. When Nichols and Bruno (2010) ask participants "What is required for some person in the future to be the same person as you?", it is possible that participants are interpreting this question in a qualitative, not numerical sense. To illustrate this, Berniunas & Dranseika draw on a convenient pair of Lithuanian phrases that disambiguate the two: *tas pats* and *toks pats* [2]. When explicitly disambiguating these two senses to their Lithuanian participants, they found that participants were significantly more likely to agree that someone was the "same person" after losing their memories, suggesting that "retention of memory may not be so critical to the preservation of individual numerical identity" (Berniunas & Dranseika, 2016, p.115).

In the context of work on the moral self effect, Starmans and Bloom (2018) similarly leverage this distinction between quantitative and qualitative identity. They claim that while Strohminger and colleagues have sought to make claims about quantitative identity ("After changing morally, can I identify *X* as the same person?"), they are actually obtaining participants' intuitions about qualitative identity ("After changing morally, is *X* dissimilar to how they were before?"). While it makes sense, they argue, that if Jim lost his moral conscience after an accident he would seem like a qualitatively different person, it wouldn't make sense to suggest that post-accident-Jim is numerically distinct from pre-accident-Jim, such that pre-accident-Jim's debts are now forgiven, or that post-accident-Jim must now get a new passport. Jim is still the same person, he's just dissimilar from before. The worry is that the measures typically used in the moral self effect studies cannot clearly differentiate between these different senses of identity. Starmans and

[2] When contrasted with toks pats, tas pats means "the same" in the sense of numerical identity while toks pats means "the same" in the sense of qualitative identity. As they reminded their participants in the studies, "If we have two white billiard balls, we can say that they are toks pats, but they are not tas pats billiard ball. If we paint one of the balls red, it is not toks pats billiard ball as before, but it is still tas pats, only painted" (Berniunas & Dranseika, 2016, p.114)

Bloom (2018) thus suggest that "we cannot tell whether these data reflect people's intuitions about similarity or about numerical identity....but we think that the most natural reading of these questions leads participants to answer in terms of similarity, not personal identity. In the real world, nobody sees moral changes as influencing identity." (p.567)

Here, we want to consider a few potential responses to these criticisms.[3] First, we are not sure that it is entirely possible to separate qualitative from quantitative identity in the way Starmans and Bloom's criticism seems to require. Second, we are not convinced that either in folk psychology or philosophy, morality is unimportant to identity.

More recent psychological continuity accounts have tended to focus on the *degree* of psychological connectedness. According to Parfit (1984), such psychological connectedness does include memories, but also psychological dispositions, attitudes, personality, preferences, and so on. This has elements of qualitative identity (a greater degree of psychological connectedness means that someone is more similar), but also numerical identity, because it is the degree of psychological connectedness that allows us to identify a person at different times as the same. Both of these points suggest that a strict division between qualitative and quantitative identity may be untenable.

There is also an empirical response to the criticisms raised above. One might think that judgements of identity change should have concomitant practical consequences. Presumably, if you judge someone to be a different person now than they were before, you would also judge that they are likely to behave differently than before. In the context of the moral self effect, a person's loss of morals should then lead participants to expect more, or worse, practical consequences for that person than with equivalent losses of memories, preferences, desires, etc. And indeed, in some of our work we have shown that compared to memories, moral changes not only affected perceptions of identity persistence (as in previous studies), but, crucially, such changes also led participants to subsequently infer a range of practical consequences, including changes in behavior, evaluations by third parties, and reductions in relationship quality (Everett, Skorburg et al., Unpublished Manuscript).

To address this question of whether moral changes are affecting judgments of (in Starmans and Bloom's terms) quantitative identity or mere similarity, we drew on the idea of special obligations: the duties we have to someone simply because of who they are. One might think, for example, that someone has obligations to their mother that they don't have to a stranger, and these special obligations towards their mother do not change even if she were to suffer a severe, debilitating illness that changed her personality. Someone's obligations and duties to her are the same because she herself *is* the same person, however dissimilar she is now to how she was in her prime. If participants judge that their own special moral obligations towards a loved one are more affected when their loved one loses their morality (compared to losing their memories), this might suggest

[3] This sections draws from and expands upon Everett, Skorburg, & Savulescu (2020).

that participants are not thinking *solely* in terms of similarity. And in fact, our results do suggest that something like this is the case (Everett, Skorburg, & Savulescu, 2020).

In our studies, we presented participants with a classic 'body switch' thought experiment in which a loved one undergoes a brain transplant with a stranger, and as a consequence either experiences no psychological change (the control condition), loses all their memories, or completely loses their moral conscience. After assessing perceptions of identity persistence, we presented a moral dilemma, asking participants to imagine that one of the patients must die (Study 1) or be left alone in a care home for the rest of their life (Study 2). In our studies, participants were made to decide who they would save or care for: the patient with their partner's brain and the stranger's body, or the patient with the stranger's brain and the patient's body.

This enabled us to test two things. First, it enabled us to replicate and extend previous empirical studies looking at whether people see psychological continuity as more important than physical continuity. And indeed, in line with previous work, in our control condition we found that participants were much more likely both to judge that the person with their partner's brain and the stranger's body was the 'real' partner, and to think that their moral duties towards this person were stronger than to the person with the stranger's brain and their partner's body. More importantly, though, by also including two conditions where, after the patient either lost all their memories or completely lost their moral conscience, we could also see whether changes to morals would be more disruptive than changes to memories for participants' perceived moral duties towards the patient.

If Starmans and Bloom's criticism (i.e. that previous studies on the moral self effect are *only* about similarity, not identity) is on the right track, then we should see no change to perceptions of moral duties depending on whether someone lost their memories or morals. The partners described in the vignettes might be perceived as more dissimilar if they have lost their morals (i.e. qualitatively different) but they would still be judged as the same person (i.e. numerically identifiable as the same person), and so presumably the special obligations would remain intact.

Indeed, we found some evidence that participants thought their moral duties towards the partner were, in fact, decreased when their partner experienced changes to their morality compared to when they experienced changes to memories, or experienced no psychological (but only physical) change. These results suggest that participants, to some extent, do perceive a person's identity to be disrupted by their loss of memories or morality, and that previous results are not only about perceived similarity (Everett, Skorburg, & Savulescu, 2020).

Taken together, the conceptual and empirical responses to challenges about the operationalization of identity concepts in the moral self literature suggests that folk intuitions about self and identity likely involve (again, using Starmans and Bloom's terms) both identity and similarity, and these are likely flexibly activated and focused on both

qualitative and quantitative identity in different contexts depending on the task at hand. We address the philosophical implications of this suggestion in Section 3 below.

2.2 Is the moral self effect generalizable?

2.2.1 Is the moral self effect only applicable to fantastical thought experiments?

For all the work that philosophers and psychologists have done on the experimental philosophy of self and identity, one might still wonder whether this is meaningful. Much of the literature (our contributions included) tends to use vignette-based studies employing far-fetched, science-fiction-like examples involving e.g., brain transplants, body-switches, magic pills, time machines, and reincarnation. While these scenarios might be helpful to clarify intuitions about philosophical thought experiments, it is also important to explore whether and to what extent the moral self effect holds in more common, everyday cases. Here again, our work has focused on the practical consequences of judgements about self and identity.

One way of testing the generalizability of the moral self effect outside the realm of thought experiments is to look at real-life cases characterized by changes to morals and memories. In this vein, Strohminger & Nichols (2015) looked at ratings of identity persistence by family members of patients with different neurodegenerative diseases. By including patients with different kinds of diseases and different symptoms, they could look at how family members judged that someone was the same person after they experienced changes to moral faculties (as in some cases of frontotemporal dementia) compared changes to memories (as in Alzheimer's disease) or changes to physical motor functions (as in amyotrophic lateral sclerosis). Mirroring the findings from thought experiments, their results show that family members of patients with frontotemporal dementia, which was associated with moral changes, rated the patient as more of a different person than did family members of patients with other kinds of diseases. Of course, while such studies provide evidence of how the moral self effect emerges in real-world contexts, their high ecological validity does come with less control: these patients will all have different presentations of symptoms, will differ in severity, and so on.

In a recent paper, we extended this line of work by taking a mixed approach using an experimental philosophy method with tightly controlled vignettes, but focusing on real-life cases of addiction (Earp, Skorburg, Everett, & Savulescu, 2019). Why addiction? A common refrain from family members and friends of addicts is that the person they knew before is not the same as the addict now. As one mother put it, "Six years have passed since I discovered that my son was using drugs. I [was] devastated, not to mention how worried I was about his well-being. My son *was not the same person anymore*" (Urzia 2014, emphasis added).

As Tobia (2017) notes, such stories are all too common: "Many have witnessed someone they loved change so profoundly that the person remaining seems an entirely

different one." Moreover, addiction is highly moralized in a way that say, dementia, is not. As a result, we hypothesized that the processes at play in the moral self effect might also arise in the context of addiction. That is, if an agent's becoming addicted to drugs leads to the perception that they are a "different person", this may be due to a presumed diminishment in moral character that such addiction stereotypically brings about. Across six studies, we found that participants judged an agent who became addicted to drugs as being closer to "a completely different person" than "completely the same person" and that these judgments of identity change are indeed driven by perceived negative changes in the moral character of drug users (Earp, Skorburg, Everett, & Savulescu, 2019).

We take these results (along with others discussed below) as evidence that the moral self effect does indeed generalize to various contexts beyond the tightly controlled, and sometimes far-fetched realm of philosophical thought experiments. We discuss some implications and potential applications in Section 3.

2.2.2 Does the direction of moral change matter?

An important question has been looming in the background so far. The moral self effect assumes that morality is at the heart of self and identity, such that changes to morals are more disruptive than other kinds of changes. This conclusion has almost exclusively been drawn from studies focusing on perceived identity persistence following either a *loss* a morals or a *loss* of memories. But what if certain features are *gained* instead?

Work on the "true self" suggests that the direction of moral change could matter. A number of studies have suggested that people typically regard others' true selves as being fundamentally good (Newman, Bloom, and Knobe 2014; De Freitas et al. 2018; Newman, De Freitas, and Knobe 2015; Bench et al. 2015). As people become more moral they are perceived to get closer to their true self or their 'essence', whereas when they become less moral, they are perceived to move further away from their true self (Bench et al. 2015; Tobia 2017).

Tobia (2015) draws on the well-worn (if potentially apocryphal) case study of Phineas Gage: a railroad worker who experienced brain damage in a horrific accident, after which he was reported to have become cruel and impulsive - so much so that "he was no longer Gage". In his work, Tobia gave participants one of two versions of this story. In one condition, participants saw the "standard" case of Phineas Gage, where he was kind before the accident, but cruel afterwards. That is, where Gage morally *deteriorated*. In another condition, holding the magnitude of the change constant, participants saw a vignette where Gage was described as cruel before the accident, but kind afterward. That is, where Gage morally *improved*. In both conditions, Tobia asked participants to judge whether Phineas Gage was the same person as before the accident. He found that Gage was less likely to be judged as identical to his pre-accident self when the change was in a "bad" direction (deteriorating from kind to cruel) than when the

change was in a “good” direction (improving from cruel to kind) even when the magnitude of the change was held constant.

These findings were further substantiated by a study demonstrating that moral enhancement is less disruptive to perceptions of identity than moral degradation and that moral degradation is especially disruptive to perceptions of identity when people expect moral enhancement (Molouki & Bartels, 2017). That said, Prinz and Nichols (2017) also report findings suggesting that whether moral changes were in a positive or negative direction did not matter: that “moral changes are regarded as threats to identity, regardless of whether those changes are good or bad” (p.454). While somewhat mixed, these findings at least raise the interesting suggestion that judgments of identity change are not solely a function of the magnitude of the change, but could be importantly related to the direction of the change. When people are perceived as deteriorating (and especially when they are perceived to deteriorate morally), they might be judged to be more of a different person than when they improve or change in a positive direction.

While this work suggests that the direction of moral change could play an important role in the moral self effect (in line with what would be predicted based on the true self literature), more work is necessary on how direction and the type of change interact - and how both of these interact depending on the target of the judgment. Perhaps, for example, gaining new memories is more disruptive than both losing memories *and* losing morals. Or perhaps all of this depends on whether participants are thinking about themselves, a friend, a stranger, or an enemy. In the same study mentioned above, Prinz and Nichols (2017) focus on judgments of the self and other, and found that the pattern for others replicated when thinking of the self: that it mattered more when the changes were moral, but it didn’t matter which direction the change were in. This, of course, goes against the suggestion in other work (e.g. Tobia, 2015, Molouki & Bartels, 2017) – perhaps different results would be obtained with a within-subjects “one change” paradigm used by Strohminger and Nichols (2016), and perhaps it matters who specifically the target is (see next section).

In recent work, we sought to shed more light on how the direction of change and target might interact. We asked participants to imagine someone changing in a variety of ways (morality, memories, warmth, and competence), where some participants read that someone increased the trait (i.e. became more moral, or gained new memories), while others, as in other studies, read that someone deteriorated (i.e. becomes less moral, or lost memories). Our results showed that changes to morality were most disruptive to perceived identity, but that the direction of change mattered too: a friend became more of a different person when they became *less* moral, but a foe became more of a different person when they became *more* moral.

Together, all these results suggest that the direction of change does matter – that even if morals tend to be more disruptive for identity than memories, losing morals can

be more impactful than gaining morals. Intriguingly, though, this also seems sensitive to the target of the thought experiment: who are we thinking of?

2.2.3 Does the target matter?

From a philosophical perspective, one might think that the primacy of morality or memories (or whatever else) for identity should be insensitive to who we're thinking about. If memories are at the core of psychological continuity, then this finding should hold for judgments about the self, but also for strangers, friends, or even enemies. And indeed, much of the work we discussed in the previous sections has found that the moral self effect does not tend to depend on the target of evaluation (e.g. first vs third person judgments). Yet these findings are somewhat surprising in light of traditional findings in social psychology that show pervasive effects of target, such that perceptions of self are often biased in certain ways in comparison to perceptions of another (and vice versa). For example, people typically think of themselves differently from others with a strong actor-observer bias (e.g. Ross, 1977) and think of those close to them differently from others (e.g. Alves, Koch, & Unkelbach, 2016; Simon, 1992), and tend to rate morality as being more important for the self and people close to them than for strangers or people they dislike (e.g. Brown, 1986; Epley & Dunning, 2000; Leach, Ellemers, & Barreto, 2007).

Given that these first and third-person asymmetries are so pervasive in the field of psychology, similar findings would be expected within the sub-field of empirical approaches to self and identity[4]. However, to date, first and third-person asymmetries have not been observed for studies investigating identity change. One of the first studies to consider the role of target showed that a body-switching paradigm yielded similar results regardless of target (Nichols & Bruno, 2010), and some of the earliest work on the moral-self effect demonstrated that moral change was more important than memories whether it was presented in a first or third-person perspective (Prinz & Nichols, 2017). More recently, no difference was observed in a series of studies directly comparing the moral self effect for self and a hypothetical other ("Chris"), albeit showing stronger effects of changes in certain moral traits on other compared to self (Heiphetz, Strohminger, & Young, 2017). Moreover, our work examining the effect of target across many different categories showed that the moral self effect holds across self, known friend, and unknown stranger, but not for a known foe (although the self and friend condition were not compared directly), again showing a small to negligible effect of target for self and (most) others (Everett, et al., Unpublished Manuscript).

It is possible that the lack of asymmetry between self and other reflects the implicit positive nature of the moral self across most targets (Strohminger, Newman, & Knobe, 2016). Indeed, the "true self" literature has demonstrated that although our own true selves are deemed to be inherently good, so are the true selves of others (Bench et al.,

[4] This section (2.2.3), and the section on neural mechanisms of target (2.3.1), are drawn from and expand upon Jordan Livingston's doctoral dissertation.

2015). This lack of the actor-observer bias for moral traits may be one reason that the effect is not observed in the studies reviewed above. However, there are other plausible avenues of explanation. Although there are well-known asymmetries for perceiving self and other, there are also asymmetries and biases for perceiving the self in the past, present, and the future (Ersner-Hershfield, Wimmer, & Knutson, 2008; Quoidbach, Gilbert, & Wilson, 2013; Bartels & Urminsky, 2011), such that thinking about the self in a hypothetical thought experiment may not be the same as thinking about the self in the here and now. Moreover, many of the moral self studies to date have not used specified targets, and using a more concrete, known target might influence the results (see Everett et al. Under Review, and Everett et al. In Prep, for exceptions). Regardless of whether the judgments change in different circumstances, the mechanisms driving the effects, or lack thereof, remain to be explored in the future.

2.3 What are the mechanisms underlying the moral self effect?

2.3.1 Neural mechanisms of target

Neuroimaging can potentially help to clarify some of the questions posed above. For example, neuroimaging studies investigating self-referential processing (in the traditional psychological sense) have found that when individuals are asked to indicate whether a series of trait words describe themselves or another individual, this processing tends to recruit the cortical midline structures of the brain (precuneus and medial prefrontal cortex), with self-referential activity recruiting activity in more ventral regions of the medial prefrontal cortex and other-referential processing recruiting activity in more dorsal regions of the medial prefrontal cortex (Denney et al., 2012; Wagner et al., 2012).

It is not entirely clear why this pattern of activity exists, and it is unlikely that the differential activity is entirely a result of the target of investigation (self or other). For example, information about the self tends to be inherently more positive than information about other individuals, and this difference is also reflected at the neural level, with self-referential neural activity sharing highly overlapping patterns with both positively-valenced information (Chavez, Heatherton, & Wagner, 2016) and value-based processing (Berkman, Livingston, & Kahn, 2017; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Information about the self, by its very nature, is also more familiar than information about other individuals, and brain regions tracking self-relevant information may be tied not only to value but also to familiarity (Lin, Horner, & Burgess, 2016).

Regardless of what is being tracked in these regions, the constellation of activity can be informative, particularly along the ventral to dorsal gradient. Indeed, studies have shown that the closeness of a target to the self can be tracked such that targets closer to the self activate more ventral regions of the medial prefrontal cortex whereas targets less close to the self activate more dorsal regions of the medial prefrontal cortex (Mitchell,

Banaji, & Macrae, 2006), with very close targets demonstrating overlapping activity (Zhu et al., 2007).

Moreover, the activity within these regions varies, to some extent, with elements of hypothetical thought. Counterfactual thinking for self and for other has demonstrated a similar ventral to dorsal gradient in the brain (De Brigard, Spreng, Mitchell, and Schacter, 2015). Thinking about the self in the future, too, tends to recruit regions of the cortical midline structures, as part of the default mode network (Buckner & Carroll, 2007; Buckner, Andrews-Hanna, & Schacter, 2008). Whereas some studies have found that thinking about the self in the future activates more dorsal (as compared to ventral) regions of the medial prefrontal cortex (Packer & Cunningham, 2009), others have found that thinking about the self in the future compared to the present simply recruits the ventral medial prefrontal cortex to a lesser degree (Ersner-Hershfield, Wimmer, & Knutson, 2008, Tamir & Mitchell, 2011; D'Argembeau et al., 2010).

This body of research offers a number of potential resources for generating hypotheses about the neural mechanisms underlying the moral self effect. At a general level, given that thinking about identity is very person-centered, activation in cortical midline structures of the brain is likely still expected. More specifically, if self-other asymmetries are absent from the moral self literature because thinking about a hypothetical self is akin to thinking about another person, overlapping activity in dorsal regions of the medial prefrontal cortex might be expected.

Alternatively, if self-other asymmetries are absent from the empirical literature on self and identity because thinking about another individual changing on a moral dimension is valuable to the self, overlapping activity in ventral regions of the medial prefrontal cortex might be expected. In any case, identifying the neural processes underlying the moral self effect could help to clarify why the primacy of morality in identity does not seem to depend on the target.

2.3.2 Neural mechanisms of values and traits

Identifying the neural mechanisms underlying the moral self effect could also help to clarify questions concerning why morality, above and beyond other personality traits, appears to be so essential to identity.

Among other things, morality can be understood as sets of norms within a society that are distinguished because they are somehow more important or more valued (see e.g., Copp, 2001), and the social nature of morality helps to highlight its value. For example, a person is likely to care about a change to her friend because she may no longer receive the benefits of being socially tied to her friend. However, the same person is also likely to care about her own morality given that it impacts her own social reputation. In both scenarios, the social features of morality ensure that changes to friends or changes to self are quite important, or valuable, to the self, and recent work supports the

idea that perceived importance to self mediates judgments of identity change (Heiphetz, Strohminger, & Young, 2017).

Incorporating evidence from neuroscience can help to assess whether the moral self effect is indeed driven by value-based processing (see May et al., this volume, for an overview of the value-based mechanisms posited by much recent work in the moral judgment and moral learning literature). Broadly, evidence increasingly suggests that moral cognition is best understood not as a unique and modular set of capacities, but rather, in terms of more domain-general processes of reasoning, value-integration, reward processing, and decision-making.

For example, Shenhav & Greene (2010, 2014) have argued that value-based processing is crucial for understanding moral judgment such that when participants are asked to imagine classic sacrificial dilemmas in which they can either do nothing and risk the death of a large group of people or do something at the expense of a single individual, value-based sub-regions of the brain are largely involved in making these types of calculations. A related set of studies have implicated value-based computation in the ventromedial prefrontal cortex during charitable giving (Hare, Camerer, Knoepfle, & Rangel, 2010; Hutcherson, Bushong, & Rangel, 2015), and yet another found that moral transgressions disrupt neural representations of value in this same region (Crockett et al., 2017). Based on this body of evidence, it seems likely that some sort of value-based processing also underpins the moral self effect such that moral traits might elicit activity in value-based regions of the brain, either selectively or more strongly than other non-moral (e.g., personality traits).

Although a value-based mechanism for the moral self effect would not be entirely surprising, it could help to push forward the field in two important ways. First, a mechanistic understanding of the moral self effect could help to elucidate the broader relationship between morality and identity. One recent commentary called into question whether morality is as important to identity, as identity is important to morality (Strohminger, 2018). We propose that the relationship between the two may be best clarified by assessing their common mechanisms.

Identity, like morality, relies on value-based processing: recent studies note strong overlap between neural activity associated with thinking about identity and value (Kim & Johnson, 2015, Northoff & Hayes, 2011), and a meta-analysis of neuroimaging studies on self and value reveals a large cluster of overlapping activation in the ventromedial prefrontal cortex (vmPFC) (Berkman, Livingston, & Kahn, 2017). Morality and identity, then, are likely intimately related via their value-based processing, and evaluating morality and identity in terms of their value-based processing may provide another currency with which to assess the nature of their relationship.

In addition to clarifying the nature of the relationship between morality and identity, a value-based approach to understanding the moral self effect could help to push forward the field of neuroscience, itself. Within the vast set of neuroscience studies that have

taken up issues of self and identity, most have utilized a task that prompts participants to rate the extent to which different personality trait words describe themselves (Denny, Kober, Wagner, & Ochsner, 2012; Wagner, Haxby, & Heatherton, 2012), an approach which remains the gold standard.

However, despite the task's popularity, very few studies have reported any neural differences in assessing the effect of trait type (Pfeifer et al., 2013; Pfeifer et al., 2009). One of the reasons why differences across traits are not reported is because most effects of traits may not become evident when using traditional univariate analysis. In contrast, more nuanced multivariate techniques are revealing neural differences in, for example, the types of information we use to organize our representations of other people (Tamir, Thornton, Contreras, Mitchell, 2016). Insights from the moral self effect highlighting the privileged status of moral (vs. non-moral) trait words and their distinct neural mechanisms (e.g., value-based processing) could motivate future neuroscience studies to continue in this tradition — using more advanced techniques to pay attention to differences between trait words.

In this sense, not only can neuroscience help to clarify mechanisms underlying the moral self effect, but insights from the moral self literature that stem from philosophy and psychology can make important contributions to the ongoing work on the neuroscience of self and identity.

3 Implications

In this final section, we will consider a few implications and applications for the moral self research program. In particular, we attempt to contextualize how the nuanced questions and concerns outlined in the previous section might have bearing on a broader range of practical issues.

3.1 Practical Implications

3.1.1 Behavior Change and Self-Regulation

One way that clarifying the psychological and neural processes involved in judgments about self and identity is important is that this research might have the potential to provide insights into translational work in domains such as self-regulation.

One example comes from the intriguing possibility that identity judgments could be leveraged as a tool for behavior change. For instance, the identity-value model of self-regulation holds that identity serves as a salient value-input for facilitating successful self-regulation and that stable, value-laden sources of identity are strongest (Berkman, Livingston, & Kahn, 2017). If morality is, as the foregoing results suggest, core to identity and is driven by value-based processing, it may be a candidate target for interventions seeking to promote behavior change. Of course, given that morality is so essential to identity, it may also be tougher to manipulate than other aspects of identity; to this effect, one recent paper found that people do not desire to be more moral, in part because they

already perceive themselves to be quite moral (Sun & Goodwin, 2019). However, just because people do not desire to be *more* moral does not mean that moral identity cannot be used as an effective motivator.

Indeed, identity has often been used as a source of moral motivation (e.g., Hardy & Carlo, 2005, 2011), and appealing to moral reasoning has been shown to motivate compliance on certain behaviors such as paying taxes (e.g., Blumenthal, Christian, Slemrod, & Smith, 2001; Ariel, 2012) and environmental conservation (Bolderdijk, Steg, Geller, Lehman, Postume, 2012; Hopper & Nielson, 1991). Given that many self-regulatory failures are often moralized (Rozin & Singh, 1999; Frank & Nagel, 2017), appealing to moral identities and values may be an effective strategy for motivating successful self-regulation, as well.

Moreover, counterfactual thought experiments, such as those traditionally used by philosophers, might also play a key role in motivating self-regulation. Many effective self-regulation techniques already draw upon hypothetical and imaginative cognitive techniques encouraging individuals to think about themselves in new and alternative ways (Kross et al., 2014; White et al., 2016). Encouraging participants to imagine the degree to which they would become a new person if they were to achieve a goal (e.g., "become a whole new you!") could provide an avenue for examining ways in which identity and value facilitate self-regulation. Exploring the mechanisms underlying the traditional moral self effect, although not directly related to translational applications, may be able to help motivate work in this direction.

3.1.2 Punishment and Responsibility

It is evident that, philosophically, identity has long been thought to be connected to moral concepts of blame, punishment, and responsibility. Practically, the research discussed in this chapter highlights how judgments about blame and responsibility are affected by perceptions of identity continuity—or disruption. We have already discussed work showing how someone's becoming addicted to drugs leads to the perception that they are a "different person", seemingly caused by perceived negative changes in the moral character of drug users (Earp et al. 2019). As another example, Gomez-Lavin & Prinz (2019) have examined the moral self effect in the context of parole decisions, finding that participants were significantly more likely to grant parole to offenders who underwent changes in their moral values, compared to mere behavioral changes. It will be interesting for future work to consider in more depth the way that appeals to identity are used in both legal and forensic settings to justify responsibility and punishment.

3.1.1 Bioethics and New Technologies

Another example which has garnered much attention as of late involves patients undergoing neurostimulation, such as Deep Brain Stimulation (DBS) treatments for

Parkinson's Disease. In one case report, after 18 months of DBS, a patient's motor symptoms were improved, but it was reported that "she was no longer able to work, had a loss of inspiration and a taste for her work and for life in general," and she said, "Now I feel like a machine, I've lost my passion. I don't recognize myself anymore" (Schüpbach et al., 2006, p. 1812). In light of this story (and many others like it), Skorburg & Sinnott-Armstrong (forthcoming) have suggested that because people tend to see moral traits as especially identity-conferring, measuring changes to moral functioning pre- and post-DBS should be a priority for neuroethicists concerned with identity changes brought about by DBS and other forms of neurostimulation.

3.2 Philosophical Implications

So far, most of our discussion has centered around experimental approaches to questions about self and identity. Here, we want to consider the extent to which such empirical work has bearing on more traditional philosophical theorizing about personal identity. Consider the following argument, adapted from Bernianus & Dranseika (2016) and Nichols & Bruno (2010):

1. If it is appropriate to consider folk intuitions in assessing theories of personal identity, then, ceteris paribus, folk intuitions that are more robust ought to be given more weight in assessing theories of personal identity.
2. Folk intuitions favoring psychological continuity accounts of personal identity are especially robust.
3. Therefore, if it is appropriate to rely on folk intuitions in assessing theories of personal identity, then, ceteris paribus, folk intuitions favoring psychological continuity accounts of personal identity ought to be given special weight in assessing theories of personal identity

Of course, one important question here is whether it is indeed appropriate to rely on folk intuitions about personal identity. After all, one might reasonably hold that what lay-people think is irrelevant to the fundamental metaphysical questions about personal identity. Perhaps the fact that ordinary people perceive morality as important to personal identity simply misses the point. What *really* matters is some metaphysical account of continuity. But this can only hold if we assume a "deep" metaphysical notion of identity, which need not be the only game in town (Prinz & Nichols, 2017). There is surely a sense in which some philosophical problems of identity are deeply metaphysical in this way. But as Prinz & Nichols (2017) point out, other questions about personal identity are not deep in the sense that they don't "depend on some hidden fact about the structure of reality". Instead, they argue, "It depends on us". We classify, label, and apply various notions of identity:

> If we are right that questions of personal identity are settled by how we do, in fact, classify, then this is a case where experimental philosophy can actually contribute to metaphysical debates. Surveys, in this case, do not just tell us what ordinary people think; they reveal the actual correlates of identity, because ordinary practices of classification determine conditions of identity." (Prinz & Nichols, 2017, p.450)

If all of this is on the right track and it is indeed appropriate to rely on the robust folk intuitions for philosophical debates, then we need to be clear on *which* folk intuitions are robust. In this chapter, we have considered numerous challenges and criticisms of the moral self effect, along with a wide range of replications and extensions. We think there is ample evidence to support the claim that the moral self effect is robust. As a result, we contend that, within psychological continuity views of personal identity, theories which emphasize the importance of morality ought to be given pride of place. Similarly, insofar as the results we have described are robust, then this could also entail that the prominent position afforded to memories within psychological continuity views may need to be revised and reconsidered. In any case, by drawing on the philosophical, psychological, and neuroscientific work on the moral self, we hope to have shown that a topic as complex and important as the moral self will surely require collaborations among philosophers, psychologists, and neuroscientists, among others.

4 References

Alves, H., Koch, A., & Unkelbach, C. (2016). My friends are all alike—the relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology*, *62*, 103-117.

Ariel, B. (2012). Deterrence and moral persuasion effects on corporate tax compliance: findings from a randomized controlled trial. *Criminology*, *50*(1), 27-69.

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.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, *33*(3), 169-185.

Bergner, R. M. (2017). What is a person? What is the self? Formulations for a science of psychology. *Journal of Theoretical and Philosophical Psychology*, *37*(2), 77.

Berkman, E. T., Livingston, J. L., & Kahn, L. E. (2017). Finding the "self" in self-regulation: The identity-value model. *Psychological Inquiry*, *28*(2-3), 77-98.

Berniūnas, R., & Dranseika, V. (2016). Folk concepts of person and identity: A response to Nichols and Bruno. *Philosophical Psychology*, *29*(1), 96-122.

Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. *Categorization inside and outside the lab*, 127-149.

Blok, S. V., Newman, G. E., & Rips, L. J. (2007). Postscript: Sorting out object persistence.

Blumenthal, M., Christian, C., Slemrod, J., & Smith, M. G. (2001). Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *National Tax Journal*, 125-138.

Bolderdijk, J. W., Steg, L., Geller, E. S., Lehman, P. K., & Postmes, T. (2013). Comparing the effectiveness of monetary versus moral motives in environmental campaigning. *Nature Climate Change*, *3*(4), 413.

Brown, J. D. (1986). Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments. *Social Cognition*, *4*(4), 353–376.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*(1), 1-38.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49-57.

Brison, S. J. (2002). *Aftermath: Violence and the Remaking of a Self*. Princeton University Press.

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2016). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex*, *27*(11), 5222-5229.

Copp, D. (2001). *Morality, normativity, and society*. Oxford University Press, USA.

Craik, F. I., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science*, *10*(1), 26-34.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879.

D'Argembeau, A., Stawarczyk, D., Majerus, S., Collette, F., Van der Linden, M., & Salmon, E. (2010). Modulation of medial prefrontal and inferior parietal cortices when thinking about past, present, and future selves. *Social Neuroscience*, *5*(2), 187–200.

De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *Neuroimage*, *109*, 12-26.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742-1752.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*, *42*, 134-160.

Dranseika, V. (2017). On the ambiguity of 'the same person'. *AJOB Neuroscience*, *8*(3), 184-186.

Earp, B. D., Skorburg, J. A., Everett, J. A., & Savulescu, J. (2019). Addiction, identity, morality. *AJOB empirical bioethics*, *10*(2), 136-153.

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction?. *Journal of Personality and Social Psychology*, *79*(6), 861.

Ersner-Hershfield, H., Wimmer, G. E., & Knutson, B. (2008). Saving for the future self: Neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience*, *4*(1), 85-92.

Everett, J. A. C., Skorburg, J. A., & Savulescu, J. (2020). The moral self and moral duties. *Philosophical Psychology*, *0*(0), 1–22.

Everett, J.A.C., Skorburg, J.A., Livingston, J.L., Chituc, V., & Crockett, M.J. (Unpublished manuscript). Morality dominates perceived identity persistence for the self, a stranger, a friend, and a foe.

Frank, L. E., & Nagel, S. K. (2017). Addiction and moralization: the role of the underlying model of addiction. *Neuroethics*, *10*(1), 129-139.

Garfield, J. L., Nichols, S., Rai, A. K., & Strohminger, N. (2015). Ego, egoism and the impact of religion on ethical experience: What a paradoxical consequence of buddhist culture tells us about moral psychology. *The Journal of Ethics*, *19*(3-4), 293-304.

Gomez-Lavin, J., & Prinz, J. (2019). Parole and the moral self: Moral change mitigates responsibility. *Journal of Moral Education*, *48*(1), 65-83.

Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: a meta-analysis. *Journal of Moral Education*, *46*(2), 97-113.

Hardy, S. A., & Carlo, G. (2005). Identity as a source of moral motivation. *Human Development*, *48*(4), 232-256.

Hardy, S. A., & Carlo, G. (2011). Moral identity: What is it, how does it develop, and is it linked to moral action?. *Child Development Perspectives*, *5*(3), 212-218.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583-590.

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology*, *78*, 210-219.

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.

Hopper, J. R., & Nielsen, J. M. (1991). Recycling as altruistic behavior: Normative and behavioral strategies to expand participation in a community recycling program. *Environment and Behavior*, *23*(2), 195-220.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451-462.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain*, *125*(8), 1808-1814.

Katzko, M. W. (2003). Unity versus multiplicity: A conceptual analysis of the term “self” and its use in personality theories. *Journal of Personality*, *71*(1), 83-114.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785-794.

Klein, S. B. (2012). The self and its brain. *Social Cognition*, *30*(4), 474-518.

Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., et al. (2014). Self-talk as a regulatory mechanism: How you do it matters. *Journal of Personality and Social Psychology*, *106*(2), 304–324.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234.

Lin, W. J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific Reports*, *6*, 28630.

Locke, J. (1975). Of Identity and Diversity. In: Essay Concerning Human Understanding. In Perry John (Ed.), *Personal Identity* (pp. 33–52). Berkeley, CA: University of California Press. (Original work published 1694).

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655-663.

Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, *93*, 1-17.

Moore, W. E. III (2015). *Sharing all the way to the bank: A neuroimaging investigation of differential self-disclosure*. Retrieved from University of Oregon Dissertation Database.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96-125.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203-216.

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

Olson, E.T. (2019). "Personal Identity", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/identity-personal/>.

Packer, D. J., & Cunningham, W. A. (2009). Neural correlates of reflection on goal states: the role of regulatory focus and temporal distance. *Social Neuroscience*, *4*(5), 412-425.

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

Paul, L. A. (2014). *Transformative experience*. OUP Oxford.

Prinz, J. J., & Nichols, S. (2017). Diachronic identity and the moral self. In *The Routledge handbook of philosophy of the social mind* (pp. 465-480). Routledge.

Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science*, *339*(6115), 96-98.

Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*(7), 485.

Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in the United States. *Journal of Consumer Psychology*, *8*(3), 321-337.

Schechtman, M. (1996). *The constitution of selves*. Cornell University Press.

Schüpbach, M., Gargiulo, M., Welter, M., Mallet, L., Béhar, C., Houeto, J. L., ... & Agid, Y. (2006). Neurosurgery in Parkinson disease: a distressed mind in a repaired body?. *Neurology*, *66*(12), 1811-1816.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741-4749.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Shoemaker, D. (2019). "Personal Identity and Ethics", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), forthcoming URL = https://plato.stanford.edu/archives/win2019/entries/identity-ethics/

Shoemaker, D. & Tobia, K. (forthcoming). Personal identity. In *Oxford Handbook of Moral Psychology.*

Simon, B. (1992). The Perception of Ingroup and Outgroup Homogeneity: Reintroducing the Intergroup Context. *European Review of Social Psychology, 3(1), 1–30.*

Skorburg, J.A. & Sinnott-Armstrong, W. (forthcoming). Some ethics of deep brain stimulation. In D. Stein & I. Singh (Eds.) *Global Mental Health and Neuroethics*. New York: Elsevier.

Starmans, C., & Bloom, P. (2018). Nothing personal: What psychologists get wrong about identity. *Trends in Cognitive Sciences*, *22*(7), 566-568.

Strohminger, N. (2018). Identity Is Essentially Moral. In K. Gray & J. Graham (Eds) *Atlas of Moral Psychology*. New York: Guildford Press

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.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*(9), 1469-1479.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159-171.

Sun, J., & Goodwin, G. (2019). Do people want to be more moral? *PsyArXiv*.

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience*, *23*(10), 2945-2955.

Tobia, K. 2015. Personal identity and the Phineas Gage effect. *Analysis 75* (3) 396–405. doi: 10.1093/analys/anv041.

Tobia, K. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics 9* (1), 37–43. doi: 10.1007/s12152-016-9248-9.

Tobia, K. (2017). Change becomes you. Aeon. September 19, https://aeon.co/essays/to-be-true-to-ones-self-means-changing-to-become-that-self .

Urzia, V. 2014. *Anthony and me*. Bloomington, IN: Xlibris.

VandenBos, G. R. (2007). *APA dictionary of psychology*. American Psychological Association.

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 451-470.

White, R. E., Prager, E. O., Schaefer, C., Kross, E., Duckworth, A. L., & Carlson, S. M. (2016). The “batman effect”: Improving perseverance in young children. *Child Development*. http://doi.org/10.1111/cdev.12695

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665.

Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage, 34*(3), 1310–1316.