Personal Identity and Dual Character Concepts[1]

Joshua Knobe
Yale University

[In press in Tobia, K. (Ed.). *Experimental Philosophy of Identity and the Self* London, Bloomsbury]

In an important recent study, Tobia (2015) gave participants a vignette about a person who gets into an accident:

> Phineas is extremely kind; he really enjoys helping people. He is also employed as a railroad worker. One day at work, a railroad explosion causes a large iron spike to fly out and into his head, and he is immediately taken for emergency surgery. The doctors manage to remove the iron spike and their patient is fortunate to survive. However, in some ways this man after the accident is remarkably different from Phineas before the accident. Phineas before the accident was extremely kind and enjoyed helping people, but the man after the accident is now extremely cruel; he even enjoys harming people.

Participants then received a simple question. Consider the man after the accident. Is that man Phineas, or would it be more accurate to say that he is not Phineas at all?

Participants answered this question on a scale that went from completely agreeing that the man is Phineas to completely agreeing that he is not Phineas. Strikingly, the mean response was at about the midpoint of the scale. In other words, people regarded this as a difficult case. They were drawn in some way toward the view that the man after the accident is Phineas, but they were also drawn in some way toward the view that he is not Phineas. Since this is an intuition about a case of radical moral change, let's refer to it as the *moral change intuition*.

In what follows, we will be looking in detail at recent empirical findings regarding the moral change intuition, but before we discuss any of those findings, it is important to see that the intuition is deeply surprising just in itself. After all, there seems to be some straightforward sense in which people think the man after the accident is *obviously* Phineas, and it's hard to see what people could possibly mean by saying that he is not.

To illustrate, suppose the man went to a bank and tried to withdraw money. People would presumably not find it remotely plausible to say: "You can't withdraw money from Phineas's account – you aren't Phineas." Similar points would no doubt hold for many of the other ordinary practices that depend on intuitions about personal identity (see Starmans & Bloom, 2018). In short, it seems that we face a puzzle about the moral change intuition. Given that there is a sense in which the man after the accident is obviously Phineas, what exactly do people mean when they say that he is not Phineas?

I will argue that if we want to understand people's intuitions in cases like this one, it will prove helpful to look to frameworks from a literature that might at first seem quite remote from the study of personal identity. Specifically, I suggest that it might be helpful to look to the literature on what are called *dual character concepts* (Del Pinal & Reuter, 2017; Flanagan & Hannikainen, in press; Guo, Dweck & Markman, in

[1] I am grateful for comments from Paul Bloom, Brian Earp, James Kirkpatrick, Shaun Nichols, John Schwenkler, Marya Schechtman and Kevin Tobia.

press; Knobe, Prasada & Newman, 2013; Leslie, 2015; Liao, Meskin & Knobe, 2020; Reuter, 2019; Tobia, Newman & Knobe, 2020).

What is a dual character concept? For a simple example, consider the criteria people ordinarily employ to determine whether or not someone counts as a scientist. Now imagine a physics professor. She spends her days running experiments and writing up papers, but she doesn't really care about getting at the truth regarding scientific questions and is dogmatically clinging to some theory in a way that ignores all of the evidence. Is this person a scientist?

In cases like this one, people tend to agree with both of the following two statements:

(1) There is a sense in which this person is a scientist.

(2) Ultimately when you think about what it really means to be a scientist, you would have to say that this person is not truly a scientist.

The fact that people agree with both of these statements provides some evidence that people actually have two different criteria for the application of this concept, hence the claim that the concept shows "dual character."

My claim will be that the frameworks developed within the study of dual character concepts give us the tools we need to understand the moral change intuition. On this view, people actually have the following pair of intuitions:

(1) There is a sense in which the man after the accident is Phineas.

(2) Ultimately, when you think about what it really means to be Phineas, you would have to say that the man after the accident is not truly Phineas.

The complicated intuition people have about the man after the accident – that he is obviously Phineas in one sense but is somehow not really Phineas in another – can then be understood as just one example of a far broader phenomenon, and it can be explained using the frameworks that have been developed to understand that phenomenon more generally.

1. A framework for understanding dual character concepts

In this first section, I introduce a general framework for understanding dual character concepts. Then, in the remaining sections, we will be looking at specific empirical results regarding the moral change intuition and using this framework to explain those results.

Within existing research, the study of dual character concepts has focused on people's use of concepts that might seem a bit distant from the moral change intuition, and I recognize that much of the material in this first section might at first appear to be irrelevant to the specific phenomenon we are trying to explain. But I promise, although many elements of this framework were originally developed to understand other concepts, every one of them will be used very directly in subsequent sections to explain the results of experiments on the moral change intuition.

Characteristic values

To begin with, let's consider the concept of *hip-hop*. Suppose we are listening to a new hip-hop song, and we start wondering about the degree to which it is continuous with the existing tradition of hip-hop music. It will be helpful here to distinguish two different ways in which we might do this.

First, we might ask whether the new song is *similar* to previous hip-hop songs. Some hip-hop songs are similar to previous songs, whereas others take things in radically new directions. This distinction might be important for various purposes.

But, importantly, we could also ask a very different question. We could ask whether the new song *embodies what hip-hop is really all about*. In this latter inquiry, we also seem to be looking at some kind of continuity with the past, but the continuity in question is of a very different type. We aren't just asking whether the new song is similar to the previous songs. Instead, we are doing something more complex. We look at the existing hip-hop songs, extract from them some deeper property ("what hip-hop is really all about"), and then ask whether the new song has that deeper property.

Importantly, this approach might yield very different judgments from the ones we would arrive at if we were just judging based on similarity. People might think that a particular song is not very similar at all to previous hip-hop songs and yet nonetheless believe that this song fully embodies what hip-hop is really all about. Or, conversely, people might think that a particular song is in most respects pretty similar to previous hip-hop songs and yet nonetheless believe that this song fails to embody what hip-hop is really all about.

One can then use this same approach when applying numerous other concepts. One can ask whether a new religious practice embodies what Christianity is all about, whether a new law embodies what the United States is all about, whether a particular scientist embodies what being a scientist is all about, and so forth. In each case, this approach enables us to think about a certain kind of continuity with the past, but in each case, that continuity is clearly not just a matter of similarity to past exemplars.

Let's now introduce some terminology that allows us to talk about these sorts of judgments. Instead of saying "what X is really all about," I will sometimes speak of the *characteristic values* of X. So I will be speaking in what follows about the characteristic values of hip-hop, the characteristic values of Christianity, and so forth.

Ultimately, our goal will be to use the notion of characteristic values to understand the moral change intuition, but first we will need to get clear in a more general way on the role of these judgments in people's cognition.

Characteristic values are not a matter of the features an object actually has

Judgments about the degree to which an object embodies the characteristic values of a concept should not be seen as just some minor variation on the idea of checking for similarity; it is a fundamentally different thing.

Suppose we want to determine whether something is similar to a given object X. We would do this by trying to figure out whether that thing has the features that X has. For present purposes, it will be important to emphasize one specific aspect of this process:

> The degree to which something is similar to object X is a matter of the degree to which it has certain features. But object X itself always has all of those features.

An obvious corollary is that nothing can ever be more similar to X than X is to itself. This is a pretty fundamental fact about the nature of similarity.

Although this might all seem pretty obvious and straightforward, it will perhaps be helpful to take a moment just to hammer home the key point. Suppose we are thinking about contemporary science (i.e., the practice of science as it exists right now). We might imagine various possible ways in which science could change, and we could ask how similar each of those possibilities would be to contemporary science. In doing so, we would be asking whether those other possibilities have certain features. However, contemporary science itself would have all of the features we were asking about. Thus, these other possibilities might be more or less similar to contemporary science, but none of them could ever be more similar to contemporary science than contemporary science is to itself.

The notion of embodying characteristic values does not work like that. To determine whether something embodies the characteristic values of X, we arrive at a judgment of the characteristic values of X, and then ask whether something embodies those values. But now notice an important fact:

> The degree to which something embodies the characteristic values of object X is a matter of the degree to which it embodies certain values. But object X might not itself perfectly embody all of those values.

A corollary is that something else might embody the characteristic values of X more perfectly than X does itself.

To illustrate, consider again the example of contemporary science. If we want to know whether something embodies the characteristic values of contemporary science, we will need to have some understanding of what the characteristic values of contemporary science are. But will we conclude that contemporary science itself perfectly embodies these values? Surely not. Thus, we might well think that some other practice – one that is a lot like contemporary science but that differs in a few specific respects – would actually more fully embody the characteristic values of contemporary science than contemporary science does itself.

If we are going to successfully make sense of this sort of judgment, we will need a conception of characteristic values according to which things can fail to perfectly embody their own characteristic values. Existing research has led to the development of a number of different theories that aim to do this. One family of theories says that people's ability to attribute characteristic values is closely tied to *teleology* (e.g., Rose, Tobia & Schaffer, 2018), while another says that it is closely tied to *psychological essentialism* (e.g., Bailey, Knobe & Newman, in press; Newman, & Knobe, 2019; Ritchie & Knobe, in press). Some recent work has sought to unify these two approaches in a theory of 'teleological essentialism' (Rose & Nichols, 2019). Nothing in what follows will depend on the resolution of this controversy, and we can therefore remain neutral as between the different possible theories.

Instead, the key point in what follows will be a relatively straightforward one that should be compatible with any plausible theory. That point is that what an object is "really all about" is not just a matter of features that the object actually has and that, as a result, an object can sometimes fail to perfectly embody the very thing that it is really all about.

When and why do people care about characteristic values?

A question now arises as to why people care about characteristic values and what role they play in people's lives. As far as I know, no existing research has tackled this question directly. I will therefore introduce a tentative hypothesis, which could be put to the test in future empirical work.

Consider a case in which people think that an object falls under a given concept. Now suppose that people conclude that the object fails to embody the characteristic values of that concept. In such cases, people will tend to think that something has gone wrong.

As an example, suppose we arrive at a conception of the characteristic values of punk rock. We might then conclude that almost all music does not embody those values. (For example, Bach's partitas don't embody the characteristic values of punk rock.) But in most cases, we would not regard this as a problem. There is no particular reason why most music ought to embody the characteristic values of punk rock, and there is therefore nothing wrong with cases in which it does not. However, something different happens when we conclude that a given piece of music actually *is* punk rock but nonetheless fails to embody the characteristic values of punk rock. In those cases, there is a mismatch, and we might feel that something has gone wrong.

This phenomenon also seems to arise for numerous other concepts. Consider the characteristic values of philosophy. Presumably, people would think that most objects do not embody these characteristic values. (Math papers, action movies and punk rock songs typically do not embody the characteristic values of philosophy.) Yet none of this is itself a problem. However, it does seem that there is a problem when a philosophy paper fails to embody the characteristic values of philosophy. In that case, there is a mismatch, and people may feel that something has gone wrong.

In introducing this idea, I mean to be making a very weak claim. The point is not that people will necessarily think that it would be better all-things-considered for every object to embody the characteristic values of the concepts it falls under. In some cases, they might think that there are also strong countervailing reasons that outweigh this one. Similarly, the point is not that people specifically think that the best way to address these mismatches is by changing the object so that it embodies the right characteristic values. In some cases, people might think it would be better to make the opposite change. For example, if a philosophy paper fails to embody the characteristic values of philosophy but does embody the characteristic values of math, people might think that the best way to address the mismatch is to turn the paper into a straight-up math paper.

With any luck, future empirical research will more directly put this claim to the test. For the moment, I adopt it as a provisional hypothesis. If it forms a part of a package that, taken together, provides an explanation for the moral change intuition, this explanation will itself provide at least some small measure of empirical support for the hypothesis itself.

Dual character statements

Consider now cases in which people represent an object as falling under a concept but think that the object fails to embody the values associated with that concept. A series of studies have explored the statements people are inclined to use in cases like this (Knobe et al., 2013).

For example, consider the hypothetical scientist we discussed in the introduction. As we noted there, participants who receive this example tend to agree with both of the following statements:

(1) There is a sense in which this person is a scientist.

(2) Ultimately when you think about what it really means to be a scientist, you would have to say that this person is not truly a scientist.

This result seems to suggest that people actually have two different sets of criteria for determining whether someone counts as a scientist. One set of criteria involves more superficial features, while the other involves characteristic values. For this reason, concepts like the concept of being a scientist are known as *dual character concepts*.

Interestingly, in existing studies on dual character concepts, it is not as though one sample of participants receives a sentence like (1) and another, completely separate sample of participants receives a sentence like (2). Rather, each individual participant receives both sentences. So participants are agreeing with both of these sentences even when they see them back to back. Indeed, participants agree even when they are conjoined to form a single sentence.

> There's a sense in which she is clearly a scientist, but ultimately, if you think about what it really means to be a scientist, you'd have to say that there is a sense in which she is not a scientist at all.

Sentences of this form have played a large role in the study of dual character concepts, and they have sometimes been referred to as "dual character statements." Studies show that people generally think that dual character statements make sense when used with dual character concepts but do not make sense when used with other concepts (Knobe et al., 2013). For example, participants think that the following sentences make sense:

> There's a sense in which she is clearly a **friend**, but ultimately, if you think about what it really means to be a **friend**, you'd have to say that there is a sense in which she is not a **friend** at all.

> There's a sense in which this is clearly a **poem**, but ultimately, if you think about what it really means to be a **poem**, you'd have to say that there is a sense in which this is not a **poem** at all.

But they think that the following sentences do not make sense:

> There's a sense in which she is clearly a **second cousin**, but ultimately, if you think about what it really means to be a **second cousin**, you'd have to say that there is a sense in which she is not a **second cousin** at all.

> There's a sense in which this is clearly a **table of contents**, but ultimately, if you think about what it really means to be a **table of contents**, you'd have to say that there is a sense in which this is not a **table of contents** at all.

Drawing on this finding, previous studies have used agreement with dual character statements as a measure of the degree to which concepts show dual character (Liao et al., 2020).

People's use of dual character statements raises a host of deep and very difficult issues. Those issues have been discussed in a number of existing papers (Del Pinal & Reuter, 2017; Guo et al., in press; Knobe et al., 2013; Leslie, 2015; Liao et al., 2020; Reuter, 2019), but I think it's fair to say that they have not yet been satisfactorily resolved. Further research on this topic is clearly needed.

In what follows, we will not be attempting to make progress on the broader questions that arise here. Instead, we will be focusing just on one very specific issue. Consider the second conjunct within a dual character statement. These are sentences like "Ultimately, she is not a scientist at all" or "Ultimately, this is not

a poem at all." We can refer to these as *ultimately-not statements*. When a person uses an ultimately-not statement, what exactly does she thereby accomplish?

Downstream effects

In thinking about this question, it will be helpful to explore some specific examples. As we will see, even a brief look at some examples suggests that things are not exactly the way we might have expected them to be if we had just considered the matter in the abstract.

Imagine that you are thinking about what is so deeply valuable in the work of Biggie, Tupac, Nas... when you are interrupted by the sound of some new song on your local hip-hop radio station. Now suppose you use an ultimately-not statement: "Ultimately, this isn't even hip-hop at all." What exactly are you doing when you use a sentence like this?

The first thing to note is that you seem to be *disparaging* the object you are discussing. Any account of these sentences that left out the disparagement would be missing something very fundamental. But the disparagement here is a complex one, and it might at first be difficult to see precisely how sentences like these can serve to disparage the objects they discuss.

To begin with, notice that it would not normally be considered disparaging to say that a song is not hip-hop. Lots of songs are not hip-hop, and there is nothing wrong with that. Of course, one might add that the sentence seems to be suggesting that the song doesn't embody the characteristic values of hip-hop, but that addition doesn't immediately address what is puzzling here. After all, lots of songs don't embody the characteristic values of hip-hop, and there is nothing wrong with that either. The disparagement in this case seems to arise from something very specific about the application of the claim to a case involving the use of a dual character concept. Let us therefore refer to it as a *dual character diss*.

Very roughly, the force of a dual character diss comes from the combination of two elements. On one hand, a certain object does fall under a concept, but on the other hand, the object does not embody those characteristic values. Thus, there is a mismatch. In some important sense the object is failing to be the very thing it actually is.

The phenomenon of dual character disses is a pervasive one, which arises in numerous different domains. Suppose you read through this paper and say: "Ultimately, this isn't even really philosophy." You would thereby be disparaging the paper, but only because there is a clear sense in which this paper *is* philosophy. Or suppose that a racist is talking about a member of some other racial group and says: "Ultimately, she isn't even really human." This is a way of disparaging that person – perhaps the worst thing that can be said about a person – but even here, one can only see that the sentence is disparaging if one understands that she actually *is* human (Phillips, 2021; cf. Smith, this volume).

In short, ultimately-not statements seem to have quite distinctive downstream effects. Consider again our ultimately-not statement:

Ultimately, this isn't even hip-hop at all.

One might at first think that the downstream effects of a statement like this one would be at least roughly similar to the downstream effects of a more straightforward statement saying that something does not fall under a concept. For example, one might think that they would be at least roughly similar to the downstream effects that would arise if someone were simply wondering whether a song was hip-hop and you answered:

No, that one isn't hip-hop.

I have been trying to suggest that this is actually not the case. The downstream effects of an ultimately-not statement aren't even roughly similar to the effects of these more straightforward statements. Rather, the downstream effects of an ultimately-not statement are closely tied to the idea of a *mismatch*. Fundamentally, what an ultimately-not statement is doing is saying that there is a mismatch between what an object is and which values it embodies.

Summary

Thus far, we have been developing a framework for understanding dual character concepts. This framework includes claims about the criteria people use in applying such concepts (characteristic values), about the linguistic expressions associated with them (dual character statements), and about the role of such linguistic expressions in people's lives (dual character disses).

The dual character framework was originally developed to understand a class of concepts that might seem quite distant from questions about personal identity, and we have been exploring that framework through a discussion of those other concepts. The key question in what follows will be whether this very same framework can also give us insight into the moral change intuition.

2. Normative standards and similarity

Admittedly, however, this is not the obvious place to go in looking for an explanation. A more natural approach would be to look to the frameworks developed within existing research on the way people think about personal identity over time. Recent studies show that these frameworks have been extraordinarily successful in explaining numerous different phenomena that seem closely related to personal identity, and one might therefore be tempted to assume that they can also be used to explain the moral change intuition.

A core goal of much of this recent research has been to understand the way people think about the various normative standards associated with being a particular person. If we determine that a particular person is me, it immediately seems to follow that this person ought to treat others in certain ways and ought to be treated in certain ways by others. If a person is me, that person has to keep my promises, teach my courses, raise my children. And, similarly, if a person is me, that person should be cared for by my friends, should be punished for my misdeeds, and so forth.

So then, how do people actually make judgments about whether a person has these normative statuses? Experimental work on this question has been deeply influenced by frameworks coming out of philosophy. Some philosophical theories suggest that personal identity is not a matter of continuity of the body but rather a matter of continuity of the mind (Locke, 1690). But, as philosophers have noted, continuity of the mind seems to be a matter of degree (Parfit, 1984). If you undergo a radical change of personality, you might be said to have a lower degree of continuity in this respect than if your personality remains pretty much constant. Experimental work has drawn on this idea in exploring the ways in which people ordinarily attribute normative standards. For example, consider a person who undergoes various changes, in the normal way, over the course of a number of years. Depending on various factors, such a person's psychology might change relatively little, or it might change quite a lot. Systematic studies show that this difference has an important effect on how people regard the agent after the change. People are less concerned with the welfare of their own future self if they believe that they will undergo very substantial psychological changes in the future (Bartels & Urminsky, 2011; Bartels, Kvaran & Nichols, 2013), and they are less inclined to punish a person for

her former misdeeds if she underwent substantial psychological changes in the time since those misdeeds (Mott, 2018).

This line of research seems to be pointing to something truly profound about the way people ordinarily understand identity over time, and it raises some fundamental, and philosophically rich, questions about the relationship between personal identity and normative status (e.g., Hershfield & Bartels, 2018; Tierney, 2020). In the present paper, however, I will not be grappling with those questions. Instead, I just want to focus in on one very specific issue. Should we be using the framework coming out of this recent research to understand the moral change intuition?

In addressing this issue, it might be helpful to start by making two simple points. First, at a very straightforward level, the thing being measured in this line of research is different from the thing being measured in studies on the moral change intuition. Studies in this line of research are concerned with the degree to which participants are inclined to help a person, or to blame a person, or to punish a person. By contrast, studies on the moral change intuition are concerned with the degree to which people agree with statements like "The man after the accident is not really Phineas." Of course, one might well think that the very same psychological processes that explain the former will also explain the latter. This is a plausible view – and I agree that it is an obvious first place to start – but all the same, it is clearly an empirical hypothesis. It might well turn out to be correct, but it is also possible that it will turn out to be wrong.

Second, in the previous section, I introduced a somewhat complex framework for thinking about certain kinds of continuity, but no one has suggested that this complex framework would be necessary for understanding the phenomena being studied in this other line of research, and my sense is that such a suggestion wouldn't even be plausible. Those other phenomena really are just a matter of *similarity*.

For example, Mott (2018) looked at judgments about punishment. The results showed that if you perform a morally bad act and then change a lot, people are disinclined to punish you for the act you performed before the change. This is a fascinating finding, but it doesn't seem at all helpful to think of it using the framework I introduced in the previous section. It is not that people are disinclined to punish you because they think that you are failing to embody what the person who performed the original act was really all about. Rather, the effect here seems to be driven by a straightforward judgment regarding similarity. People are disinclined to punish you because you are now so different from the way you were at the time you performed the act.

In sum, existing research on intuitions about personal identity has uncovered some very surprising effects on people's judgments about normative standards, and those judgments appear to be driven by perceived similarity. What we want to know now is whether the frameworks that have proven so helpful in explaining these effects will help us in understanding the moral change intuition.

3. The moral change intuition

To address this question, let's now turn to existing empirical work on the moral change intuition. We will be focusing on four major findings.

Moral vs. non-moral

Strohminger and Nichols (2014) reported a series of studies that looked at which specific changes led to the intuition that personal identity was disrupted. For example, in one study, all participants were asked to imagine a person named Jim who suffers a brain injury in a car accident. They were then randomly assigned

to be told that one specific aspect of Jim's previous mental states was lost as a result: his memories, his desires, his perceptual abilities, or his moral conscience. Participants were then asked about the degree to which they agreed with the sentence: "The transplant recipient is still Jim."

In a striking result, Strohminger and Nichols found that a loss of moral conscience led to a different pattern of responses than any of the other sorts ofpsychological changes. In all other conditions, participants tended on the whole to agree with this sentence, but in the condition where Jim loses his moral conscience, they tended on the whole to disagree.

This same finding also emerged in numerous further studies, including everything from studies in which participants are asked to imagine that one person's soul enters another person's body to studies on people whose spouses actually are undergoing psychological changes due to dementia (Heiphetz, Strohminger & Young, 2017; Prinz & Nichols, 2016; Strohminger & Nichols, 2014; Strohminger & Nichols, 2015). There seems to be a pervasive tendency whereby people's use of these sentences is much more affected by changes in moral traits than by other sorts of changes.

This is not the pattern we would have expected to find if we thought that these intuitions were driven by straightforward judgments of similarity. After all, similarity does not seem to be just a matter of having the same moral traits; it seems to be a matter of having the same features more generally (same memories, same preferences, etc.). We might be able to accommodate this result on a theory that emphasizes similarity, but to do so, we would have to introduce some additional complexities into our account of the similarity judgment that plays a role here. For example, we could say that it isn't a matter of similarity across the board but rather a matter of similarity in one specific respect (e.g., similarity with regard to moral traits).

Good vs. bad

Thus far, we have seen that intuitions about personal identity depend especially on changes in moral traits, but does it matter whether those changes involve morally good traits or morally bad traits?

To address this question, Tobia (2015) conducted an elegant experiment. Participants were randomly assigned to receive either a vignette in which Phineas loses morally good traits or morally bad traits. In the condition in which the morality good traits are lost, participants received the vignette quoted at the beginning of this paper. In the condition in which morally bad traits are lost, participants received a modified version [changes in boldface type] :

> Phineas is extremely **cruel**; he really enjoys **harming** people. He is also employed as a railroad worker. One day at work, a railroad explosion causes a large iron spike to fly out and into his head, and he is immediately taken for emergency surgery. The doctors manage to remove the iron spike and their patient is fortunate to survive. However, in some ways this man after the accident is remarkably different from Phineas before the accident. Phineas before the accident was extremely **cruel** and enjoyed **harming** people, but the man after the accident is now extremely **kind**; he even enjoys **helping** people.

In both conditions, participants were asked whether they thought that the man after the accident was not Phineas.

The results showed a surprising asymmetry. In the condition where the morally good traits are lost, participants' ratings were at about the midpoint of the scale (indicating that they were uncertain whether to say that the man after the accident was Phineas). By contrast, in the condition where the morally bad traits

were lost, participants were specifically inclined to say that personal identity was not disrupted and that the man after the accident was still Phineas.

Subsequent studies found similar effects with other materials (Earp, Skorburg, Everett & Savulescu, 2019), across a wide variety of cultures and languages (Dranseika et al., unpublished data), and even in studies with children (Lefebvre & Krettenauer, 2020). For example, in a study that we will be discussing further below (Tobia, 2015; based on a case from Parfit, 1984), participants were given a vignette about a Russian nobleman. In one condition, participants were told that he started out with egalitarian ideals and then lost those ideals (morally good traits lost). In the other condition, participants were told that he started out with empty egalitarian ideals and then lost those ideals (morally bad traits lost). Participants were more inclined to say that the person after the loss of ideals was not the same as the original nobleman in the condition where morally good traits were lost.

This pattern in the results looks even farther from the pattern one would expect if these intuitions were driven by similarity judgments. To hold onto the idea that these intuitions are driven by similarity judgments, we would therefore have to introduce even more complexity into account of the similarity judgments themselves. Not only would we need to say that they are a matter of similarity in one specific respect, we would have to say that the notion of similarity at work here is asymmetric. Notice the structure of the studies we are discussing. In both conditions, the person is described as undergoing a change between the very same two states (being morally good, being morally bad), and the only difference is the direction in which the person is moving (good to bad vs. bad to good). Thus, to explain this result in terms of similarity, we would have to say that being morally good is seen as more similar to being morally bad than being morally bad is seen as similar to being morally good.

At this point, it is beginning to seem that the concept of similarity is not actually playing any helpful role in explaining the results. What the studies show is that people are especially inclined to use a certain sort of sentence in cases where a person fails to show a morally good trait that he or she showed previously. We might be able to rig up a very complex account of similarity that allowed us to accommodate this result, but is the concept of similarity actually helping us in any way to make sense of it?

Beyond human beings

So far, we have been looking at intuitions about human beings and their psychological states. A question arises, however, as to whether these same effects would emerge if one looked at other types of objects.

In a series of studies, De Freitas and colleagues therefore took the effect that Tobia found for intuitions about Phineas and asked whether that same effect would emerge for intuitions about non-human objects (De Freitas, Tobia, Newman & Knobe, 2017). For example, in one study, participants were asked to imagine a physics paper called "Atom Dynamics." They were told that the authors revised the paper, deleting some of the existing sections and adding some new sections. In one condition, participants were told that these changes involved eliminating all of the good parts of the paper, while in the other condition, participants were told that these changes involve eliminating all of the bad sections of the paper. All participants were then asked whether the paper after the changes genuinely was "Atom Dynamics." The results showed the same effect observed for people's intuitions about persons. Participants were more inclined to say that the paper was no longer really "Atom Dynamics" when it lost its *good* properties. This effect also arose for nonhuman objects of many other types: a university, a nation, a conference, a rock band.

In short, when we look specifically at intuitions about human beings, we find a difference between losing good traits and losing bad traits. But this effect does not appear to be due to something unique about human beings. Rather, it seems to be just one instance of a far more general effect involving a difference between losing good properties and losing bad properties.

Downstream effects

We have seen that people are more inclined to have a certain sort of intuition when a person loses morally good traits then when that person loses morally bad traits. A question now arises about the downstream effects of this intuition. If people aremore inclined to think that the man after the accident is not really Phineas when he loses morally good traits, what impact will this have on the way they actually think about or interact with him?

Work on the role of similarity has emphasized one specific type of downstream effect. Specifically, this work points to an impact of similarity judgments on intuitions about normative standards. Does the moral change intuition work in that very same way? If people are more inclined to think that the man after the accident is not truly Phineas when he loses morally good traits, will they be less inclined to think that the normative standards that applied to the original Phineas still apply after he loses morally good traits?

Although existing studies have not explored this question using the Phineas case in particular, Earp and colleagues explored this question by looking at intuitions about advance directives (Earp, Latham & Tobia, 2020). Suppose Robin signs an official document saying that if she ever ends up in a certain kind of medical condition, she does not want to be resuscitated. Subsequently, she undergoes radical psychological change as a result of dementia, and she either loses her morally good traits or her morally bad traits. Then the person who exists after the psychological change gets pneumonia and ends up in precisely the medical condition described by the advance directive. Should the doctors' treatment of the person after the onset of dementia be governed by the advance directive that the original Robin signed?

The question being asked here has very much the same structure as the questions asked in the many existing studies on personal identity and normative standards. There is a normative standard that applies to a person before a change (the advanced directive), followed by a change in that person's psychological states (the onset of dementia). Participants are then asked a question design to see whether they continue to apply the normative standard even after this change. The key question is whether participants will be less inclined to apply the normative standard when the change involves the loss of morally good traits than when the change involves the loss of morally bad traits. Strikingly, the results showed that people were *not* less inclined to apply the normative standard when the change involved a loss of morally good traits. That is, there was no effect at all such that participants were less inclined to think that the doctors should obey the advance directive in the condition where morally good traits were lost than in the condition where morally bad traits were lost.

Because this study came out only very recently, I worry that the field might not yet have absorbed its full significance. Previous studies consistently find that people are more inclined to have a certain sort of intuition when morally good traits are lost than when the morally bad traits are lost, and a question now arises about the downstream effects of that intuition. The obvious first hypothesis would be that it has exactly the same sorts of downstream effects that have been demonstrated in numerous existing studies on the effects of similarity judgments. But we are now getting some evidence that it does *not* have those same sorts of downstream effects. This provides at least some reason to suspect that it is not the same thing but is something else entirely.

Summary

Our aim here has been to understand the intuition people express when they use sentences like: "The man after the accident isn't even really Phineas." One initial question we face is whether this intuition can be straightforwardly explained using the sorts of frameworks that come out of the existing philosophical literature on personal identity or whether we will need a new sort of framework to understand it.

To make progress on this question, we reviewed four findings from the existing empirical literature: (1) People's use of these sentences is especially affected by changes in moral traits and (2) even more so by the loss of morally good traits. (3) This same basic pattern arises for the way people describe things other than human beings. (4) Although the loss of morally good traits has an especially large impact on use of these sentences, it does not have an especially large impact on people's intuition about the normative standards that apply to the person after the loss. None of these findings seems to follow in any obvious way from the frameworks that have been so successful in helping us understand the relationship between normative standards and psychological similarity.

So then, how are the findings to be explained? One possible approach would be to try fiddling with the frameworks developed within this prior work. We might then end up with something that resembled those frameworks but that also differed from them in various details. For example, we might say that people have some very complex notion of similarity that is actually asymmetric. Or we might say that, for some complex reason, intuitions about personal identity don't have the impact one might think they would on judgments about advance directives.

Although there is a chance that this strategy will ultimately prove successful, I will be pursuing a very different approach. On the view I will be developing, these findings do not provide any reason to revise existing frameworks. Those frameworks are fine just as they are; the issue is simply that the moral change intuition is completely unrelated to the phenomena they were originally designed to explain. Thus, if we want to explain the moral change intuition, we will have to switch over to an entirely different approach.

4. The moral change intuition and dual character concepts

Let's therefore shift gears and ask instead whether we can make sense of these results within the framework introduced to understand dual character concepts. On the hypothesis I will be proposing, it is not just that there is some loose analogy between the moral change intuition and dual character concepts. Rather, the moral change intuition simply *is* an example of the use of a dual character concept.

On this view, then, what people are doing when they say "The man after the accident isn't really Phineas" is deeply similar to what people are doing when they say things like "Ultimately, this song isn't really hip-hop." The best way of understanding the sentence about Phineas is through an application of the framework that was first developed to understand those other cases.

Now, of course, these two cases look very different from a metaphysical perspective. In the former case, we have a concept that applies to various different *objects*. We are now thinking about a particular object, and we are wondering whether this concept applies to it. By contrast, in the latter case, we have a concept that applies to various different time slices of a person, i.e, to what metaphysicians sometimes call *person-stages*. We are now thinking about a particular person-stage, and we are wondering whether the concept applies to it. Clearly, there is an important difference between thinking about objects and thinking about person-stages.

Given this obvious difference, it would certainly be reasonable to suspect that it won't be possible to find a single overarching framework that helps make sense of both sorts of cases. Nonetheless, I willbe

arguing that the dual character framework actually does capture a more abstract structure that applies across both. To see how this might work, we need to engage in a more detailed examination of the moral change intuition itself.

Dual character statements

As we noted above, one way to test whether a concept has dual character is to see whether people are willing to use it in sentences of a specific form that we have called "dual character statements." To see whether people are inclined to express the moral change intuition using sentences of this form, I conducted a simple experiment.

Two hundred and five participants were recruited using Amazon's Mechanical Turk. All participants received one of Tobia's (2015) vignettes about Phineas. Participants in one condition received the version in which Phineas loses morally good traits; participants in the other condition received the version in which Phineas loses morally bad traits. All participants then received the dual character statement:

> 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.

Participants rated this statement on a scale from 1 ('completely disagree') to 7 ('completely agree').

Ratings in the condition where the morally good traits were lost (M = 5.5, SD = 1.4) were significantly higher than were ratings in the condition where the morally bad traits were lost (M = 4.6, SD = 1.7), $t(204)=3.7$, $p < .001$. (All data and R code for this study are available at https://osf.io/w842x/)

To get a better understanding of these results, it might be helpful to compare them to the results of the original Tobia study. Figure 1 shows the original Tobia results; Figure 2 shows the present results.

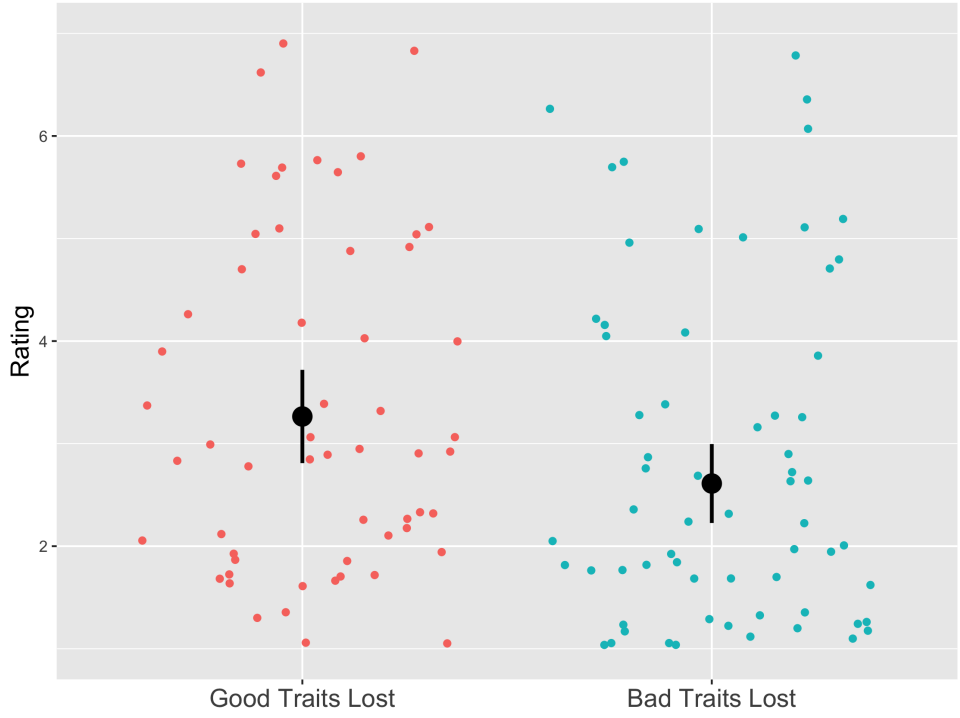

Figure 1. Jittered plot showing the results of Tobia (2015) . Each colored point represents the rating given by one individual participant. Black circles show the means for each condition. Error bars show 95% CI.

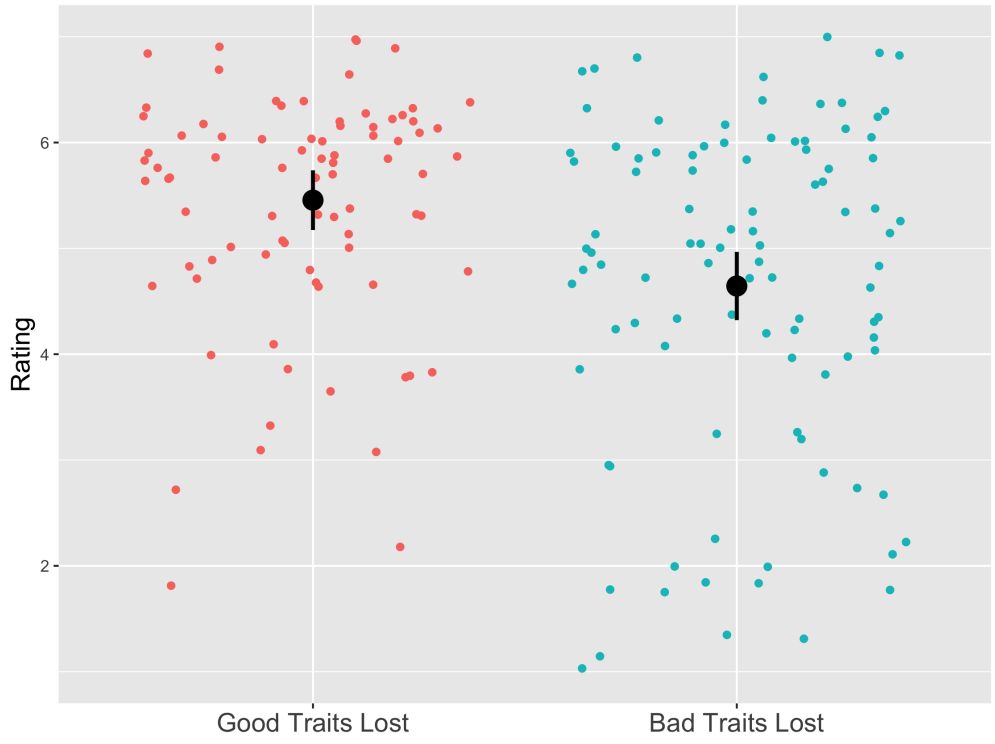

Figure 2. Jittered plot showing the results of the present study. Each colored point represents the rating given by one individual participant. Black circles show the means for each condition. Error bars show 95% CI.

This simple study yields two findings that we will be trying to explain.

First, in the condition where morally good traits were lost, participants tended on the whole to agree with the dual character statement. This result contrasts with the result from the original Tobia study. In that original study, participants showed at least some willingness to say that the man after the accident was not truly Phineas, but it was not the case that they actually tended on the whole to agree with that statement. What we find here is that when participants are given the more complex dual character statement, the majority actually *agree*. The obvious conclusion would be that this more complex statement more accurately captures their opinions about this case.

Second, the results from the condition in which morally bad traits were lost were very different from the results in the conditions in which morally good traits were lost. In the condition in which morally bad traits were lost, participants' judgments were all over the place. Some agreed with the dual character statement, others disagreed. Thus, an adequate account of people's judgments in these cases needs to explain

why this second condition is different from the first and why it leads to so much disagreement in people's intuitions.

In sum, the present study provides at least some evidence for the hypothesis that the moral change intuition is best understood in terms of dual character concepts. The key question now is whether we can use the dual character framework to explain the exact pattern of intuitions that people show in these cases.

Criteria

Consider the original Phineas, as he existed before the accident. In the actual text of the vignettes participants received, there is a certain amount of information about what the original Phineas was actually like, and this information makes it clear that he was morally good in one condition, morally bad in the other. But to apply the dual character framework, we need to look at something that goes beyond just what Phineas was actually like. We need to look at *what he was really all about*. Importantly, these two things might sometimes differ substantially; what Phineas was actually like might involve a failure to embody the characteristic values that constituted what he was really all about.

What will participants conclude in each of these conditions about what Phineas was really all about? This sort of question has been explored in a number of recent studies, and we now have at least some amount of evidence regarding the answer. When an agent is described as having morally good features, people tend to think that the agent is fundamentally morally good. By contrast, however, when an agent is described as having morally bad features, people don't just conclude that the agent is fundamentally morally bad. Instead, people seem to regard this as a difficult or confusing case. Some people say that the agent is morally bad, while others say that there is some sense in which, deep down, the agent is morally good (Newman, De Freitas & Knobe, 2015).

Of course, further questions immediately arise as to why people think about morally bad agents in this way. These questions have been a major preoccupation of recent work in this area. The tendency people show in these cases seems to be related to a more general tendency to believe that, deep down, all agents are morally good (Newman, Bloom & Knobe, 2014; Strohminger, Knobe & Newman, 2017), which in turn seems to be related to an even more general tendency to think that all objects, including non-human objects, have good essences (e.g., De Freitas, Tobia, Newman & Knobe, 2017). The question as to how to explain this more general tendency is a difficult one, and I don't have anything new to contribute to it here. The key point for present purposes is just that if people's judgments do show this pattern, we can use those judgments to explain the moral change intuition.

First, consider the condition in which Phineas loses his morally good traits. In that condition, people tend to think that the man after the accident is failing to embody the characteristic values of the original Phineas. For this reason, they say that, ultimately, the man after the accident is not truly Phineas.

Now consider the condition where Phineas loses his morally bad traits. In that condition, the man after the accident is certainly very dissimilar from the original Phineas, so if these intuitions were driven by similarity judgments, people should again say that the man after the accident is not Phineas. However, on the view we have been developing, these intuitions are *not* driven by similarity judgments. Instead, they are driven by judgments about characteristic values.

This gives us a very different way of explaining the results in that condition. Given the way in which people attribute characteristic values, it seems likely that people will be all over the place when it comes to judgments about what the original Phineas was really all about. Some people will think that the cruelty he shows on the surface is what he is really all about, while others will think that despite the features he shows

on the surface, what he was really all about was something more morally good. These different people should have different intuitions about whether the man after the accident embodies the characteristic values of the original Phineas. Those people who think that what the original Phineas was all about was something morally bad should conclude that the man after the accident does not embody the characteristic values of the original Phineas, but those participants who think that what the original Phineas was all about was something morally good should reach the opposite conclusion. They should conclude that the man after the accident embodies the characteristic values of the original Phineas even *more* than the original Phineas himself did.

Downstream effects

When one first encounters the moral change intuition, it is only natural to seek to understand it in terms of a familiar web of concerns involving normativestandards. After all, if we learn that a particular person is Phineas, we immediately begin to attribute to that person certain normative standards. He has to fulfill Phineas's promises, he has the right to use Phineas's possessions, he should be punished for Phineas's misdeeds, and so forth. If someone says that the man after the accident is not truly Phineas, an obvious first thought is that this person is saying that some of these normative standards do not fully apply to the man after the accident.

I have been arguing that the moral change intuition should not be understood in this way. It doesn't have anything to do with any of the concerns that are familiar from the existing literature on personal identity. Rather, it should be understood as closely connected to the use of ultimately-not sentences with dual character concepts.

As soon as one begins thinking in this way, a completely different set of concerns immediately suggest themselves. Suppose we think about Phineas and thereby extract a view about what he is all about. We could then pick out any person and ask whether that person embodies Phineas's characteristic values. For example, we could ask whether Barack Obama embodies Phineas's characteristic values, or whether Alexander the Great embodies those values. Yet, though we could ask this question about any arbitrary person, we usually would not care very much about the answer. However, there is one specific person who has a special relationship to Phineas's characteristic values. That person is *Phineas*. To the extent that Phineas fails to embody what Phineas is all about, it seems that something is going wrong. He is failing to *be himself*.

This point comes out even more clearly when we consider more ordinary cases. Take the case of the Russian nobleman (Parfit, 1984). Suppose you knew the Russian nobleman back when he passionately believed in the cause of liberating the serfs. Then you see him again, years later, and he seems interested only in preserving his own power and privilege. You say: "It isn't even really *him* anymore." What exactly would you be conveying with a sentence like that?

On the hypothesis we are exploring here, you would be pointing to a certain sort of mismatch. The person you are seeing is clearly still the person you once knew, but at the same time, he is tragically failing to embody the values of the person you once knew. Therein lies the force of the claim you are making about him – that he is failing to be the very thing that he is.

5. Conclusion

Our inquiry has been concerned with the relationship between two things. On one hand, there are very general theories about dual character concepts. On the other, there are a series of specific empirical findings concerning the moral change intuition. We have been exploring the hypothesis that the former can explain the latter.

To evaluate this hypothesis, we have been looking at some theories regarding dual character concepts in general and at some findings regarding the moral change intuition in particular. If we consider just the findings that are already available, it does seem that the theories provide good explanations for the findings. This gives us at least some reason to suspect that our hypothesis might be on the right track.

But of course, the hypothesis makes predictions that go far beyond anything that can be verified just by looking at existing findings. In the years to come, we will presumably uncover further facts both about dual character concepts in general and about the moral change intuition in particular. The hypothesis we have been exploring makes a prediction about those further facts. It predicts that even after we know much more about the dual character concepts and about moral change intuition, we will continue to find an explanatory relationship between the two.

References

Bailey, A., Knobe, J. & Newman, G.E. (in press). Value-based Essentialism: How Beliefs About Shared Values Promote Essentialist Beliefs. *Journal of ExperimentalPsychology: General.*

Bartels, D. & Urminsky, O. 2011. On Intertemporal Selfishness: The Perceived Instability of Identity Underlies Impatient Consumption. *Journal of ConsumerResearch* 39: 182-198.

Bartels, D.l M., Kvaran, T. & Nichols, S. 2013. Selfless giving. *Cognition* 129: 392-403.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science*, 41, 382-402.

Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, 41, 477–501.

Dranseika V., Lauraitytė E., & Experimental Jurisprudence Cross-Cultural Study Swap Consortium (unpublished data). Cross-cultural Replication of Tobia (2016).

Earp, B. D., Latham, S. R., & Tobia, K. P. (in press). Personal transformation and advance directives: an experimental bioethics approach. *American Journal of Bioethics*

Earp, B. D., Skorburg, J. A., Everett, J. A. C., & Savulescu, J. (2019). Addiction, identity, morality. *AJOB: Empirical Bioethics*, 10, 136–153.

Flanagan, B. & Hannikainen, I. (in press). The Folk Concept of Law: Law is Intrinsically Moral. *Australasian Journal of Philosophy*

Guo, C., Dweck, C. S., & Markman, E. M. (in press). Gender categories as dual-character concepts? *Cognitive Science*.

Heiphetz, L, Strohminger, N. & Young, L. 2017. The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science* 41: 744- 767.

Hershfield, H. E., & Bartels, D. M. (2018). The future self. The psychology of thinking about the future, 89-109.

Knobe, J., Prasada, S., & Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127, 242–257.

Lefebvre, J. P., & Krettenauer, T. (2020). Is the true self truly moral? Identity intuitions across domains of sociomoral reasoning and age. *Journal of ExperimentalChild Psychology,* 192, 104769.

Leslie, S.-J. (2015). "Hillary Clinton is the only man in the Obama administration": Dual character concepts, generics, and gender. *Analytic Philosophy*, 56(2), 111–141.

Liao, S. Y., Meskin, A., & Knobe, J. (2020). Dual character art concepts. *Pacific Philosophical Quarterly*, 101(1), 102-128.

Mott, C. 2018. "Statutes of limitations and personal identity." In T. Lombrozo, J. Knobe & S. Nichols (eds.) *Oxford Studies in Experimental Philosophy* 2: 243-269.

Newman, G. E., & Knobe, J. (2019). The essence of essentialism. *Mind & Language*, 34, 585-605.

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

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

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Phillips, B. (2021). 'They're Not True Humans': Beliefs About Moral Character Drive Categorical Denials of Humanity. Unpublished manuscript. https://psyarxiv.com/5bgxy

Prinz, J., & Nichols, S. 2016. "Diachronic identity and the moral self." In Julian Kiverstein (ed.), *The Routledge Handbook of Philosophy of the Social Mind*Abingdon: Routledge: 449-464.Reid,

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, 14(1), e12557.

Ritchie, K & Knobe, J. (in press). Kindhood and essentialism: Evidence from language. *Advances in Child Development and Behavior* (Ed.) M. Rhodes.

Rose, D., Tobia, K. & Schaffer, J. 2018. Folk teleology drives persistence judgments. *Synthese*.

Rose, D. & Nichols, S. (2019), Teleological Essentialism. *Cognitive Science*, 43: 12725.

Smith, D. L. (this volume). "Human" Is an Essentially Political Category. In Tobia, K. (Ed.). Experimental Philosophy of Identity and the Self, London, Bloomsbury.

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

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

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

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.

Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. (2014). How many of us are there? *Advances in Experimental Philosophy of Mind*, 181.

Tierney, H. (2020). The subscript view. A distinct view of distinct selves. *Oxford Studies in Experimental Philosophy* Volume 3, 3, 126.

Tobia, K. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75(3), 396–405.

Tobia, K. P., Newman, G. E., & Knobe, J. (2020). Water is and is not H_2O. *Mind & Language*, 35(2), 183-208.