CellPress
REVIEWS

Letter

Moral Goodness Is the Essence of Personal Identity

Julian De Freitas,[1,*] Mina Cikara,[1] Igor Grossmann,[2] and Rebecca Schlegel[3]

Starmans and Bloom ([1]; henceforth S&B) argue that research on the centrality of morality in people's intuitions about personal identity does not reveal much about people's notions of personal identity (whether an individual is the same person at time$_a$ and time$_{a+1}$), but only something about their notions of similarity (how much the person at time$_a$ shares properties with the person at time$_{a+1}$). We agree with S&B that it is important to distinguish between these constructs but disagree with their conclusion. Here we briefly review evidence that judgments regarding personal identity following a change in moral character cannot be explained by a mere (dis) similarity account.

First, consider their thought experiment: 'Suppose that when Bob was 20, he was the nicest of people. Generous, kind to animals – a real mensch. But then Bob experienced a profound moral transformation, and he turned into a terrible person: mean, selfish, psychopathic, a man who robs stores and kicks dogs.' Now, invert their thought experiment: 'Suppose that when Bob was 20, he was a terrible person: mean, selfish, psychopathic, a man who robs stores and kicks dogs. But then Bob experienced a profound moral transformation, and he turned into the nicest of people. Generous, kind to animals – a real mensch.' Comparing this moral improvement scenario to the moral deterioration scenario offered by S&B, we see that the magnitude of change is the same (i.e., dissimilarity is equated). Nonetheless, we and others consistently find that in such improvement scenarios people say that Bob is still the same person (reviewed in [2]). Why? Because people believe that Bob's true self, his essence, consists of the morally good traits (e.g., [3,4]). Thus, when Bob undergoes significant moral improvement, people believe that his good essence has emerged, that he has 'found himself'. But when Bob undergoes significant moral deterioration, people believe that he is no longer there. Furthermore, we find via mediation analyses that participants' judgments of similarity between such cases do not explain identity judgments, whereas beliefs about a morally good essence do [5].

Second, we disagree that people in these experiments are merely speaking colloquially when they agree that the person is no longer present. When Phineas Gage's head was penetrated by a tampering iron, his family said he was 'no longer Gage' [6]. Presumably they were not just casually indicating that he was no longer a nice guy, but really did feel that the person they knew and loved was no longer present. The same goes for patients who have undergone significant neurodegeneration [7]. We think that their relatives are saying more than just that their loved one has undergone a big psychological change. Rather, they are picking out something deeper: despite there still being a superficial physical presence, this seems to be just a shell that is no longer inhabited by the person who they came to know and love.

Third, regarding passports, birthdays, and taxes, we think it is important to distinguish between personal identity and legal identity [8]. Even if people have the sense that someone is no longer the same person, the superficial characteristics of the person pertaining to their legal identity remain the same. At the very least, there continues to be a body with the same perceptual abilities, occupying the same spatiotemporally continuous path, with the same indisputable physical birthdate and history. So in one superficial sense, people are forced to admit, as a Capgras patient might even admit, that the 'person' is still there. But in another sense – the sense that counts for personal identity – people cannot shake the feeling that the person they know is gone.

Finally, people's beliefs about a good true self appear to be a form of psychological essentialism (PE) (reviewed in [2]), which has previously been linked to category membership and identity using precisely the sorts of studies that S&B recommend (e.g., [9,10]). An important property of PE is that removing an entity's seemingly essential characteristics is more disruptive to its identity than removing its seemingly peripheral characteristics. This is exactly the pattern observed in recent experiments: removing morally good traits leads to a larger sense of disruption to personal identity compared with other kinds of traits, including morally bad traits of an equal magnitude. Furthermore, beliefs about a good true self show various hallmarks of PE [11]. People believe that morally good traits are innate and cross-temporally stable, that there is a boundary separating the self-essence from other aspects of the self, and that self-essences have non-obvious properties and are diagnostic of what is true about an individual. Finally, like other documented effects of PE, the good true self belief seems to operate similarly across cultural and individual differences [12]. To our minds, the most parsimonious interpretation of these various findings is that people believe that moral goodness is the fundamental quality that defines the person. Eliminate this quality, and you eliminate the person.

[1]Department of Psychology, Harvard University, Cambridge, MA, USA
[2]Department of Psychology, University of Waterloo, Waterloo, ON, Canada
[3]Department of Psychology, Texas A&M University, College Station, TX, USA

*Correspondence: defreitas@g.harvard.edu (J. De Freitas).
https://doi.org/10.1016/j.tics.2018.05.006

References
1. Starmans, C. and Bloom, P. (2018) Nothing personal: what psychologists get wrong about identity. *Trends Cogn. Sci.* 22, 566–568
2. De Freitas, J. *et al.* (2017) Origins of the belief in morally good true selves. *Trends Cogn. Sci.* 21, 534–636
3. Newman, G.E. *et al.* (2014) Value judgments and the true self. *Pers. Soc. Psychol. Bull.* 40, 203–216
4. Newman, G.E. *et al.* (2015) Beliefs about the true self explain asymmetries based on moral judgment. *Cogn. Sci.* 39, 96–125
5. De Freitas, J. *et al.* (2017) Normative judgments and individual essence. *Cogn. Sci.* 41, 382–402
6. Harlow, J. (1868) Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society* 2, 327–347
7. Strohminger, N. and Nichols, S. (2015) Neurodegeneration and identity. *Psychol. Sci.* 26, 1469–1479
8. Tobia, K.P. (2015) Personal identity and the Phineas Gage effect. *Analysis* 75, 396–405
9. Gelman, S.A. (2003) *The Essential Child: Origins of Essentialism in Everyday Thought,* Oxford University Press
10. Gutheil, G. and Rosengren, K.S. (1996) A rose by any other name: preschoolers' understanding of individual identity across name and appearance changes. *Br. J. Dev. Psychol.* 14, 477–498
11. Christy, A.G. *et al.* (2017) The essence of the individual: the pervasive belief in the true self is an instance of psychological essentialism. *PsyArXiv* Published online October 24, 2017. http://dx.doi.org/10.17605/OSF.IO/K3JBA
12. De Freitas, J. *et al.* (2017) Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cogn. Sci.* 42 (Suppl. 1), 134–160

Letter

If You Become Evil, Do You Die?

Christina Starmans[1,*] and Paul Bloom[2]

De Freitas *et al*. [1] agree with us about the importance of distinguishing between personal identity and similarity. We agree with them that individuals can be obliterated through severe neurodegeneration and the like: as we put it in our original article [2], 'There are cases . . . where it may be thought that a person ceases to exist while their body survives, as in severe dementia.' [3]. Finally, we share an appreciation of research on the notion of a morally good true self (indeed, one of us is a coauthor on the first paper on the topic).

However, our disagreement is a major one. We think that when someone becomes immoral, people see them as undergoing a substantial transformation. But the person does not cease to exist; we think the answer to the question 'If you become evil, do you die?', is plainly 'no'. The arguments made by De Freitas *et al*. suggest that they think that the answer is 'yes'. They argue that moral goodness is seen as intrinsic to a person's existence: 'Eliminate that quality, and you eliminate the person'. If Bob loses his goodness, they say, then 'people believe he is no longer there'.

For them, this follows from the true self findings they review [4–6]. This research explores what people see as the most important, essential, or central features of an individual. But we see this as being distinct from the project of determining the features that lead individuals to persist over time. There is a world of difference between thinking that the most important, essential, or central feature of Bob is his kindness, and thinking that if Bob were to lose his kindness, he would cease to exist.

De Freitas *et al*. point out that true self research reveals an asymmetry – it is seen as more of a change when someone goes from good to bad than when someone goes from bad to good [4–7]. We agree, and indeed we are happy to endorse their own account of this, which is that when someone becomes good, people believe that this reflects the influence of an already existing true self and is thus less of a transformation. However, none of this is a challenge to our view that these transformations preserve personal identity.

On a minor point, De Freitas *et al*. also insist that we take literally the language people use to describe dramatic moral changes. They argue that when Gage was described as 'no longer Gage', 'presumably [Gage's family] were not just casually indicating that he was no longer a nice guy'. We agree that there is nothing casual here about this type of statement – brain damage is serious business – and presumably the family meant something broader, akin to 'Gage no longer has the important traits that we've always associated with him'. But we see this as similar to saying, 'I'm just not myself today', which obviously cannot be meant literally, and illustrates that we typically use this type of language to talk about changes, not obliteration.

Finally, we agree that it is important to distinguish between personal and legal identity; there are cases where a legal notion (ownership, consent, culpability) is different from the psychological one. However, legal identity is at least partially based on intuitions about personal identity. As a science fiction example, imagine that Bob dies and his body is donated to science. Fred, sound of mind but poor of body, has his brain transplanted into Bob's body. In this case we would certainly assign the person who looks like Bob – the body that used to be Bob's – a new name and a new legal identity. Why then do we not do this when a person becomes immoral? We believe, in this type of case, that nobody thinks there is a numerically different person. Instead, the perception is that there is one person who has changed dramatically.

We conclude with some street corner experimental philosophy, asking our readers this: have you ever encountered someone, either in real life or in fiction, who started off good and then become immoral? If so, did the person then disappear? Did their body become a shell, now occupied by a different individual? We take it that the answer is 'no'. De Freitas *et al*. might say that we are being unfair – they are not saying that individuals who become immoral literally cease to exist. But we do not know how else to interpret their strong claims, such as 'eliminate the person' and 'he is no longer