The good ship Theseus: The effect of valence on object identity judgments

Julian De Freitas (julian.defreitas@psy.ox.ac.uk)
Department of Psychology, University of Oxford. 9 South Parks Road, Oxford OX1 3UD, United Kingdom

Kevin Tobia (kevin.tobia@philosophy.ox.ac.uk)
Department of Philosophy, University of Oxford. Radcliffe Humanities, Woodstock Road, Oxford OX2 6GG, United Kingdom

George E. Newman (george.newman@yale.edu)
School of Management, Yale University. New Haven, CT, 06511, USA

Joshua Knobe (joshua.knobe@yale.edu)
Program in Cognitive Science, Yale University. New Haven, CT, 06511, USA

Abstract

How do ordinary people decide whether an individual object at t_1 is the *same* individual at t_2? We show that *valence*—people's value judgments about whether a given trait is good or bad—can influence this decision. This effect is explained by people's tendency to believe that the underlying essence of an entity is good, and may be part of a far wider phenomenon of how people understand essences in general—be they of humans, categories, or even non-human objects.

Keywords: Concepts; moral reasoning; identity; persistence; essentialism; true self

Would the USA still be the USA without freedom of expression? What about without social discrimination? Would the Beatles still be the Beatles without John Lennon? What about without Pete Best?

How do people decide that an individual object at t_1 is the *same* individual at t_2? While there have been several theories proposed in metaphysics about what *should* constitute identity (in a normative sense), here our focus is descriptive—in other words, what are the ways in which everyday people tend to make judgments of identity continuity? A good deal of philosophical and empirical work has investigated the lay theories people use when judging whether an individual object continues to be the same individual (see Rips, Blok, & Newman, 2006). Much of this work has converged on four factors: *phenomenalism* (whether the object maintains the same appearance across time or transformation), *sortalism* (whether the object maintains the same basic-level category membership, e.g. 'dog'), *physicalism* (whether the object continues to be composed of the same physical 'stuff'; e.g. the 'ship of Theseus' is a thought experiment that questions whether an object that has all its parts replaced remains the same object), and the *causal continuer view* (whether the object has the same underlying cause).

However, to date research on identity judgment has not explored the role of *valence*, i.e. of people's value judgments about whether certain traits are good or bad. At first, it might seem strange to even ask whether valence can influence something as concrete as whether an object is the same individual. Yet a recent wave of research has suggested that valence—at least, in the form of moral valence—influences people's judgments about all sorts of matters which initially seem to have nothing to do with value judgments (see Knobe 2010), including *freedom* (Phillips & Knobe, 2009), *weakness of will* (May & Holton, 2012), *intentional action* (Knobe, 2003), and *happiness* (Phillips, Misenheimer, & Knobe, 2011). We predict that valence has a similar effect on people's identity judgments.

Valence and Essence Ascriptions

More specifically, we propose that the impact of valence on identity judgments is explained by people's *psychological essentialism* (Gelman & Hirschfeld, 1999; Keil, 1989; Medin & Ortony, 1989). In its everyday sense, an essence is often referred to as something intrinsic to an entity that makes it the kind of thing that it is (Newman & Keil, 2008). Much of the existing work on essentialism has been concerned with category essences (e.g. Gelman, 2003; Keil, 1989), but people also ascribe essences to individual entities (e.g. Gupta, 1980; Gutheil & Rosengren, 1996). As an example, the essence of the present paper is constituted in part by its engagement with questions about identity, and if we eliminated all discussion of these questions from the paper, its very essence would have been removed. By contrast, this paper happens to start with the letter 'W', but that is not the essence of the paper, and if we changed that one letter, the essence of the paper could still remain.

Recent research suggests that valence may have an impact on intuitions about individual essence. In particular, studies of the way people understand human beings indicate that people are more inclined to say that the *good* qualities of a human being constitute that human being's essence (Newman, Bloom, & Knobe, 2014; Newman, De Freitas, & Knobe, in press). In fact, even if a human being behaves immorally, people will still be inclined to say that the human being is good 'deep down' (Newman et al., 2014; Newman et al., in press).

Our hypothesis is that this very same effect arises when people think about things other than human beings. That is,

if you are thinking about entities such as a band, or a nation, or even inanimate objects such as a science paper, you will also be inclined to see the good aspects of it more as being its essence. Since people should be more inclined to see an entity as losing its essence when it loses good traits than when it loses bad traits, they should be more inclined to say that the entity itself no longer even exists when its good traits disappear than when its bad traits disappear. In short, we predict that the impact of valence on essence judgments leads to an impact on intuitions about object identity.

The Present Studies

To test this hypothesis, we chose five different non-human entities and described them as either improving or deteriorating along a relevant dimension. Studies 1 and 2 tested whether there would be an asymmetry in people's judgments about whether the object before and after the changes was the same individual. Studies 3 and 4 tested whether these results could be explained in terms of people's psychological essentialism.

Study 1

Study 1 investigated whether people show an asymmetry in persistence judgments depending on whether an entity undergoes a valence change to become more positive (Improvement), or negative (Deterioration). We predicted that if an entity improves, people will believe that the entity after the changes is still the same entity as the entity before the changes, whereas if the entity deteriorates, people will believe that the entity after the changes is no longer the same entity as the entity before the changes.

Methods

320 participants (M_{age} = 30, 104 female) were recruited using Amazon's Mechanical Turk. Participants were assigned to one of ten conditions in a 2 (valence: improvement vs. deterioration) X 5 (vignette) design. The different vignettes served merely as a robustness check, and included a band, science paper, nation, university, and conference. For example:

Eastford is a large university. When the university first opened, some of its departments used diverse teaching styles and also challenged students to think for themselves, while others taught by reading straight out of a textbook and did not allow any student participation. Over the years, some of the original departments were removed, and some new departments created.

Now after these changes, the majority of departments teach by using diverse teaching styles and also challenging students to think for themselves.

OR:

Now after these changes, the majority of departments teach by reading straight out of a textbook and do not allow any student participation.

Participants were then asked to rate the extent to which they agreed with the statement (1 = 'completely disagree', 7 = 'completely agree'): *The new [Eastford] is not really the same [university] as the [Eastford] before the changes*. Finally, they answered two multiple choice comprehension questions about the vignette: *Before the changes, how would you describe [Eastford]?* and *After the changes, how would you describe [Eastford]?* Participants chose an answer from the same three options for both questions, for example: a) Used diverse teaching styles and also challenged students to think for themselves, b) Taught by reading straight out of a textbook and did not allow any student participation, and c) Some departments used diverse teaching styles and also challenged students to think for themselves, while others taught by reading straight out of a textbook and did not allow any student participation

Results and Discussion

83 participants were excluded for not answering all comprehension questions correctly. A 2 (valence: improvement vs. deterioration) X 5 (vignette) ANOVA indicated a significant main effect of valence on persistence judgments, $F(1,227) = 30.24$, $p < .001$, $\eta_p^2 = .118$. As predicted, participants were significantly more likely to agree that the entity after the changes was no longer the same as the entity before the changes when it deteriorated (M= 5.33, SD= 1.39) than when it improved (M= 4.35, SD= 1.43). There was also a main effect of vignette, in which participants gave higher persistence ratings to some vignettes, $F(4,227) = 5.24$, $p < .001$, $\eta_p^2 = .085$, though critically this factor did not interact with the primary variable of valence, $p = .982$, and all vignettes were directionally consistent with our hypothesis (i.e. a higher likelihood of reporting identity disruption when the entity deteriorated than when it improved). In short, participants were significantly more inclined to say that the entity's identity was preserved when it improved than when it deteriorated.

Study 2

Results from Study 1 provided initial support for asymmetric identity judgments of non-human entities based on valence, but another possible explanation of these results is that they were driven by judged asymmetries in *intentionality* and/or the *quantity of good and bad traits* of the entity. For instance, perhaps participants were more inclined to think that the good traits of the entity were intended by its creator than to think that the bad traits were intended in this way. Then it might have been this inference about the creator's intentions, rather than anything about valence directly, that was driving the effect. Along similar lines, participants may have judged an entity to have *more* good than bad traits before it changed. Then, perhaps they judged the good traits to be more essential to the entity simply because good traits happened to be more prevalent in this specific entity.

To address these alternative explanations, Study 2 included explicit information about intentionality (we also tweaked the original vignettes wherever we thought intentionality was potentially ambiguous), and described the conditional change as going from a *majority* good (bad) to a *majority* bad (good). Doing so ensured that there would be no ambiguity in participants' minds that a full valence change had taken place in each of the conditions, since now it would be very clear what the initial and final valences were. Finally, in order to strengthen our confidence that the previous effect observed in Study 1 was truly one on *identity* judgments per se (and not some related, but vaguer notion), we also added a second, more direct measure of persistence.

Methods

320 participants (M_{age} = 30, 96 female) were recruited using Amazon's Mechanical Turk. The experimental design was largely consistent with Study 1, except this time information about intentionality was explicitly added, and entities were described as changing from a majority good (bad) to a majority bad (good). For example:

Bellshore is a small country. In the majority of its regions the local government intentionally teaches people to express their opinions freely in public, while in some other regions the local government intentionally teaches people to discriminate against one another for being different.

Over the years, some regions of Bellshore change their policies. Now, after these changes, in the majority of regions the local government intentionally teaches people to discriminate against one another for being different.

OR:

Bellshore is a small country. In the majority of its regions the local government intentionally teaches people to discriminate against one another for being different, while in some other regions the local government intentionally teaches people to express their opinions freely in public.

Over the years, some regions of Bellshore change their policies. Now, after these changes, in the majority of regions the local government intentionally teaches people to express their opinions freely in public.

Participants were then asked to rate the extent to which they agreed with the statement (1 = 'completely disagree', 7 = 'completely agree'): *[Bellshore] after the changes is not really the same [country] as [Bellshore] before the changes.* The order of this question was counterbalanced with a second question about persistence:

Person A and Person B agree that at a superficial level (e.g. the number of regions, or the number of people) Bellshore before the changes shares a lot in common with Bellshore after the changes. However, when they consider what it really means to be Bellshore, the country, they run into a disagreement about what has happened to the identity of Bellshore after the changes:

Person A thinks that Bellshore after the changes is still the same country as Bellshore before the changes.
Person B thinks that it makes more sense to say that Bellshore is no longer the same country it used to be. The way he sees it, the original Bellshore no longer exists.

Who do you agree with more, Person A or Person B?

Participants answered this question on a 7-point Likert Scale (1 = 'person A', 4 = 'equally agree with both persons', 7 = 'person B'). Finally they also answered two comprehension questions about the vignette.

Results and Discussion

86 participants were excluded for not answering all comprehension questions correctly. The two items measuring persistence intuitions showed high internal consistency ($\alpha = 0.85$), and thus were averaged to produce a single measure of identity judgment. A 2 (valence: improvement vs. deterioration) X 5 (vignette) ANOVA indicated a significant main effect of valence on identity judgments, $F(1,224) = 19.24$, $p < .001$, $\eta_p^2 = .079$. Participants were significantly more likely to agree that the entity after the changes was no longer the same as the entity before the changes when it deteriorated (M=5.67, SD=1.18), than when it improved (M=4.99, SD=1.38). There was also a main effect of vignette, in which participants gave higher identity ratings to some vignettes, $F(4,224) = 4.86$, $p < .01$, $\eta_p^2 = .080$, though critically this factor did not interact with the primary variable of valence, $p = .301$, and all vignettes were directionally consistent with our hypothesis (i.e. a higher likelihood of reporting identity disruption when the entity deteriorated than when it improved). Thus, these results replicate the original effect via two different measures of identity, while simultaneously ruling out alternative explanations based on potential, intuited differences in intentionality or the initial quantity of a valenced trait.

Study 3

Study 3 tested whether this asymmetric effect of valence on identity judgments can be explained by participants' intuitions about the individual essences of these entities.

Methods

320 participants (M_{age} = 28, 91 female) were recruited using Amazon's Mechanical Turk. The experimental design was almost identical to that of Study 2, though this time we used only the second identity measure from Study 2. Following this measure, participants rated their agreement with the statement, *[Entity] after the changes no longer reflects the true essence of the original [Entity]* (1 = 'completely disagree', 7 = 'completely agree'). Previous studies have successfully employed this kind of wording to probe intuitions about essentialism (e.g. Newman et al., 2014; Newman et al., in press). Finally, participants answered two

comprehension questions about the vignette.

Results and Discussion

89 participants were excluded for not answering all comprehension questions correctly. A 2 (valence: improvement vs. deterioration) X 5 (vignette) ANOVA indicated a significant main effect of valence on identity judgments, $F(1,221) = 19.67$, $p < .001$, $\eta_p^2 = .082$. Participants were significantly more likely to agree that the entity after the changes was no longer the same as the entity before the changes when it deteriorated (M=5.53, SD=1.43) than when it improved (M=4.57, SD=1.70). There was no main effect of vignette, $p = .085$, and this factor did not interact with the primary variable of valence, $p = .430$.

To determine whether beliefs about essence explain the effect of valence on identity ratings, we then conducted a bootstrap mediation analysis (Preacher & Hayes, 2008; Hayes, 2012), with condition as the independent variable, ratings of identity persistence as the dependent variable, and measures of essence as a potential mediator. The analysis indicated that essence did indeed significantly mediate the effect of valence on identity judgments (95% CI = -1.19 to -.56; see Figure 1).

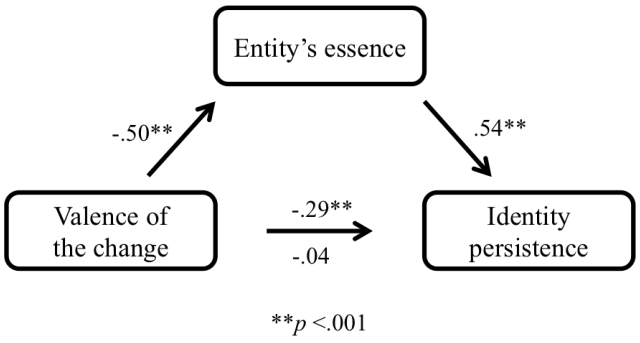

Figure 1: Mediation results from Study 3.

Study 4

Study 3 provided evidence that the asymmetric effect of valence on identity judgments can be explained by people's intuitions about the essence of these entities: since positively valued traits are seen as part of the very essence an entity[1], manipulating whether these traits are present leads to a corresponding influence on people's intuitions about identity persistence. But in order to really demonstrate that essence is *causally* responsible for the current effects, one would need to directly manipulate the essence of an entity and show that this leads to a corresponding change in people's intuitions about whether the identity of the entity still exists.

Study 4 took exactly this approach. We reasoned that even though people assume by default that the essence of an entity is good, if they are directly told that the essence of an entity is bad then this should lead to a corresponding 'flip' in their intuitions about whether the identity of the entity is still the same after the changes. In particular, a bad-essenced entity that deteriorates should now be seen as staying in line with its true bad essence, and so people should be more inclined to say its identity is the same after the changes; by contrast, a bad-essenced entity that improves should now be seen as deviating from its true bad essence, and so people should be more inclined to say its identity is not the same after the changes.

Note that such a study is an extremely direct test of our hypothesis that intuitions about essence explain the present asymmetries. We are predicting that, even if one only manipulates whether the entity's essence is initially described as good or bad (i.e. while keeping all other details of the vignette identical for all participants), this information about the essence will causally determine whether people think its identity survives the changes.

Study 4 would also rule out another possible explanation of our results, which argues that the only reason people are more inclined to say an entity's identity no longer exists after it deteriorates, is to communicate their disapproval of the negative resulting characteristics. Note that our prediction for the current study is that if an entity is described as having a bad essence, then people should be more inclined to say exactly the opposite: that its identity no longer exists after it *improves*.

Methods

640 participants (M_{age} = 31, 210 female) were recruited using Amazon's Mechanical Turk, and assigned to one of four conditions in a 2 (essence: good vs. bad) X 2 (valence: improvement vs. deterioration) design. Participants initially read one of two essence stems about a band:

Ever since the band Breath String was formed, it was clear that there was something distinctive about its music. It sometimes played bad [good] songs, but deep at its essence it was a fundamentally authentic [superficial] band. At the very core of its existence, the band was never only [only ever] interested in playing songs if it thought that doing so would make it famous and rich, and it had

[1] Of course people do not always value the same things, and can sometimes even have very different values about the same thing. This observation naturally opens us to another prediction: when people value opposing characteristics in the same entity, they should exhibit correspondingly different views about whether a valenced change to that characteristic constitutes an improvement or deterioration, which should in turn influence their intuitions about its identity. In a separate study we showed exactly this effect, by comparing liberals vs. conservatives' intuitions about entities that changed toward having either more characteristics valued by liberals or more characteristics valued by conservatives, e.g. a conference whose presentations now almost only dealt with climate change vs. military defense technology. As predicted, liberals were more likely to say that the entity's identity was lost when it changed toward conservativism, while conservatives were more likely to say that the entity's identity was lost when it changed toward liberalism.

great interest [no interest at all] in making good quality music.

In the early 2000s, the band then went through a transitional phase. It was very confused about many things in its repertoire and the members regularly abused drugs and alcohol. Most of the bands that Breath String would perform alongside were basically playing low [high] quality and unoriginal [original] music, but at its essence there was something that made Breath String fundamentally different from all of them.

On the next page, you will read about that period in Breath String's history, and then how things turned out in the very end.

As a manipulation check, participants then used a 7-point Likert Scale (1 = fundamentally bad, 7 = fundamentally good) to answer the following question: *Based on this information, how would you characterize Breath String's "true essence"?*

On a separate page, participants were then shown a similar vignette as used in Study 2, in which the band was described as changing either toward playing only superficial and commercial songs (deterioration), or only moving and meaningful songs (improvement). Then all participants read the description of two people disagreeing about whether *Breathstring* was still the same band after the changes (see Study 2 for exact wording), and indicated which of the two people they agreed with more by using a 7-point Likert Scale (1 = 'person A', 4 = 'equally agree with both persons', 7 = 'person B'). Lastly, they answered two comprehension questions about the vignette.

Results and Discussion

197 participants were excluded for not answering all comprehension questions correctly. Results from the manipulation check question indicated that we successfully manipulated essence judgments: participants were significantly more likely to rate the band as good when they read the 'good' stem (M=4.84, SD=1.39) than when they read the 'bad' stem (M=2.57, SD=1.30), $t(441)=17.72$, $p < .001$.

Turning to the main dependent variable, a 2 (essence: good vs. bad) X 2 (valence: improvement vs. deterioration) ANOVA indicated a significant interaction between essence and valence, $F(1,439) = 9.08$, $p < .01$, $\eta_p^2 = .02$. Consistent with our hypothesis, when the entity was described as having a good essence participants were significantly more likely to agree that the entity after the changes was no longer the same after it deteriorated (M=5.65, SD=1.41) than after it improved (M=5.12, SD=1.46), $t(222)= 2.80$, $p < .01$. Conversely, when the entity was described as having a bad essence participants were more likely to agree that the entity after the changes was no longer the same after it improved (M=5.61, SD=1.14) than after it deteriorated (M=5.32, SD=1.66), although this difference did not reach statistical significance, $t(217)= -1.47$, $p = .143$.

This last result identifies a boundary condition for the present effects that is in agreement with Studies 1-3, suggesting that people find it most natural to assume the essence of entities is good. In other words, even when people were explicitly told that the essence of the entity is bad, the identity effects were attenuated compared to the effects for the good essence stem (although the effects for both stems were still in the predicted directions). This suggests that, unless people are given explicit information to believe otherwise, they by default assume that the essence of an entity is good. One notable exception to this trend may be rare cases of extremely harmful entities, such as concentration camps, or terrorist groups, to which people might naturally posit a bad essence. Exploring people's intuitions about such entities remains an intriguing avenue for future research, and may be informed by the current studies.

General Discussion

Across four studies, we found that people are more inclined to say that a non-human entity no longer exists when its good traits disappear than when its bad traits disappear. Study 1 demonstrated this basic influence of valence on intuitions about object identity, Study 2 replicated the effect while controlling for potential confounds, and Studies 3 and 4 showed that this phenomenon can be explained in terms of people's beliefs about the *essence* of an entity. The studies also cumulatively suggest that people by default assume that the essence of an entity is good.

Although the present studies were limited to five entities, we see no reason to believe that these intuitions would not extend to numerous others. Indeed, the fact that we find this intuition for such a diverse group of entities suggests that there are likely all sorts of other entities that frequently give rise to this same intuition. An interesting question for future work will be to explore the precise characteristics of entities that allow them to be 'essentialized' in this way and whether there are entities for which the current intuitions would not hold.

One central question that has arisen from work on the essence of humans is why it is that people see this essence as being good in the first place (Newman et al., 2014; Newman et al., in press). One possibility is that this effect arises from a tendency to think of humans, per se, in positive terms (Sears, 1983). After all, moral traits are a useful predictor of how people will fare as cooperative partners in all sorts of social interactions (e.g. Nowak, Page, & Sigmund, 2000), and so (the theory goes) it would make sense for these traits to have a substantial influence on identity judgments. Indeed, as some have pointed out, morality may even be the one capacity that is uniquely human— "even *C. elegans* has memory" (Strohminger & Nichols, 2014). But what all of these theories have in common is that they assume the explanation lies with

humans. That is, they assume that we already know what the boundaries of the problem are, and all we need to do now is continue researching humans and eventually we will discover the correct explanation.

The present results suggest that this assumption is mistaken. Instead, people's default tendency to see the essence of humans as good actually appears to be part of a far wider phenomenon that can be found in people's way of understanding essences in general—be they of humans, categories, or even non-human objects.

Acknowledgments

We would like to thank Christina Starmans and Jillian Jordan for useful suggestions.

References

Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology*. Cambridge, MA: MIT Press.

Gupta, A. (1980). *The logic of common nouns: An investigation in quantified modal logic*. New Haven: Yale University Press.

Gutheil, G., & Rosengren, K. S. (1996). A rose by any other name: Preschoolers understanding of individual identity across name and appearance changes. *British Journal of Developmental Psychology, 14,* 477–498.

Hayes, A. F. (2012) PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [white paper].

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33,* 315–329.

May, J. & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*, 341–360.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, England: Cambridge University Press.

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

Newman, G. E., De Freitas, J., & Knobe, J. (in press). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science.*

Newman, G. E., & Keil, F. C. (2008). Where is the essence? Developmental shifts in children's beliefs about internal features. *Child Development, 79,* 1344–1356.

Nowak, M., Page, K., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*, 1773–1775.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, *20*, 30–36.

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review*, *71*, 929–937.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Rips, L. J., Blok, S., & Newman, G. (2006). Tracking the identity of objects. *Psychological Review*, *113,* 1–30.

Sears, D. O. (1983). The person positivity bias. *Journal of Personality and Social Psychology, 44,* 233–250.

Strohminger, N. and Nichols, S. (2014). The essential moral self. *Cognition*.