Journal of Experimental Social Psychology 93 (2021) 104100

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The true "me"—Mind or body?

Iris Berent *, Melanie Platt

Northeastern University, United States of America

ARTICLE INFO

Keywords:
True self
Free will
Dualism
Essentialism
Core knowledge
Intuitive psychology

ABSTRACT

Laypeople construe one's life narrative around a single protagonist – the true self. Who is this true self? Does it reside in our mind or body? Is it only aligned with one's biological essence, or also with their moral core, the home of free will? In three experiments, participants reasoned about John—a modern-day reincarnation of Dr. Jekyll and Mr. Hyde. John's character was evaluated by two tests (brain and behavioral), whose outcomes diverged (e.g., a brain test indicating benevolence; a behavioral test indicating aggression). Results showed that participants aligned John's free will with his good acts (irrespective of test), but they defined his essence by the outcomes of the brain test. We interpret the results to suggest that people hold conflicting tacit notions of the true self. One's freely-willed moral core is good, but one's essence is aligned with the body.

1. Introduction

Reflecting on a person's life, people can often recall acts of kindness and unkindness; generosity and selfishness. Despite these conflicting attributes, however, people believe that, deep down, there is a single, unitary protagonist—the "true self" (De Freitas, Cikara, Grossmann, & Schlegel, 2017; Newman, Bloom, & Knobe, 2014a; Strohminger, Knobe, & Newman, 2017). Who is this unique, unitary "me"?

The existing literature has defined the "true self" in terms of two attributes. On the one hand, the "true self" is defined in moral terms, and specifically, as good (De Freitas, Cikara, et al., 2017; Newman, De Freitas, & Knobe, 2015; Strohminger et al., 2017); on the other hand, the true self is aligned with one's (biological) essence (e.g., Heiphetz, 2019). These two attributes have been assumed to be seamlessly intertwined (e.g., De Freitas, Cikara, et al., 2017; De Freitas, Tobia, Newman, & Knobe, 2017; Heiphetz, 2019; Strohminger et al., 2017). Here, however, we show that they stand in sharp tension.

The tension arises because morality is appraised in terms of free will—a notion that is typically perceived as disembodied (e.g., Greene & Cohen, 2004; Nichols, 2011). Biological essence, however, is firmly anchored in the body (e.g., Newman & Keil, 2008). The following investigation thus explores who the "true me" is —is it aligned with my essence or my moral core? My mind or my body?

1.1. My true self: my good moral essence

Several recent proposals suggest that people hold a notion of the "true self", distinct from the self. While the self is aligned with a wide range of psychological attributes (e.g., memories, personality, intelligence), the "true self" is primarily a moral notion (De Freitas, Cikara, et al., 2017; Newman et al., 2015; Strohminger et al., 2017). For example, participants believe that, if some of a person's attributes were to change (e.g., as a result of swallowing a pill), then that person would be less likely to remain "the same" if the change affected their moral core compared to changes affecting their personality, perceptual abilities, memories, and desires (Heiphetz, Strohminger, & Young, 2017; Strohminger & Nichols, 2014). This suggests that the person's core—their true self—is perceived as defining who they are, and as moral in nature.

Moreover, people perceive the true moral self as good (De Freitas, Cikara, et al., 2017; Newman et al., 2015; Strohminger et al., 2017). When a protagonist exhibits an abrupt change to their moral fiber (e.g., a corrupt policeman turning honest vs. an honest policeman turning corrupt), participants are more likely to identify the protagonist's true self with their better acts (De Freitas & Cikara, 2018; Molouki & Bartels, 2017; Newman et al., 2014a; Tobia, 2016). What counts as "good" can vary—liberals believe that the true self is revealed when a sexist person is transformed into a libertarian; for conservatives, it's the transformation of an unpatriotic person into a patriot that unveils one's true self (Newman et al., 2014a). But within each group, people believe that the traits they consider to be good are likely to persist (Newman et al.,

* Corresponding author at: Department of Psychology, Northeastern University, 125 Nightingale Hall, 360 Huntington Ave., Boston, MA 02115, United States of America.

E-mail address: i.berent@neu.edu (I. Berent).

https://doi.org/10.1016/j.jesp.2020.104100
Received 16 October 2020; Received in revised form 15 December 2020; Accepted 22 December 2020
Available online 27 January 2021

2014a). Similar belief in the good true self obtains across countries (the US, Russia, Singapore, and Colombia; De Freitas et al., 2018).

These observations, then, would seem to suggest that the true self is perceived not only as morally good but also as immutable—people believe that the true self remains invariant throughout life (De Freitas, Cikara, et al., 2017). Accordingly, several researchers have attributed the belief in the underlying true self to psychological essentialism (e.g., De Freitas & Cikara, 2018; De Freitas, Cikara, et al., 2017; Heiphetz, 2019; Newman, Bloom, & Knobe, 2014b; Strohminger et al., 2017; Strohminger & Nichols, 2014).

Essentialism is the intuitive belief that living things are what they are because they possess some inborn, immutable essence (Gelman, 2003; Keil, 1986; Medin & Ortony, 1989). Children, for example, believe that offspring maintain properties of their biological parents (Gelman & Wellman, 1991; Hirschfeld, 1995; Solomon, Johnson, Zaitchik, & Carey, 1996), and that its essence is immutable—a raccoon does not turn into a skunk by painting its fur (Keil, 1986).

Beliefs about the true self bear these hallmarks of essentialism. Indeed, children and adults believe that a person's goodness is inborn and immutable, and these essentialist beliefs are stronger when people consider the goodness of a person compared to their badness (Heiphetz, 2019).

Summarizing, then, the existing literature has captured the perceptions of the true self in two ways. On the one hand, the true self defines one's moral core, and it is specifically good; on the other, it is aligned with one's inborn immutable biological essence. These two attributes—that of (good) moral core and (biological) essence—would seem to harmoniously coexist (e.g., De Freitas, Cikara, et al., 2017; De Freitas, Tobia, et al., 2017; Heiphetz, 2019; Strohminger et al., 2017). In this view, then, one's essence is one's good, moral core. As we next see, however, these two notions conflict.

1.2. My "True Me": Mind or Body?

The tension between "my good moral core" and "my essence" arises because, in intuitive psychology, the notions of "morality" and "essence" contrast with respect to their perceived anchoring in the body. To appreciate this tension, however, we must first consider how the notion of morality is related to yet another construct—that of free will.

Indeed, moral acts are typically identified as acts that are committed freely (e.g., Nichols, 2011; Nichols & Knobe, 2007; Roskies & Nichols, 2008; Sarkissian et al., 2010) and intentionally (Barrett et al., 2016; Greene & Cohen, 2004; Nichols, 2011); actions committed under duress (e.g., a gun is pointed to one's head) are not perceived as reflecting on one's moral character. If the true self is indeed defined in moral terms, then acts committed freely should be perceived as more indicative of one's true self. In line with this possibility, past results suggest that, when participants' belief in their free will is diminished, they are reportedly less aware of their true self, and they are also less likely to consider their moral decisions (e.g., donation to charity) as reflective of their authentic true self (Seto & Hicks, 2016). Whether one's true self is indeed perceived as the source of one's free will is uncertain—the results from this single study are insufficient to settle this question. But given that the true self is firmly linked to morality (specifically, to goodness), and morality, in turn, is typically predicated on free will, the possibility that the true self is endowed with free will seems like a plausible hypothesis. Fig. 1 captures this hypothesis graphically (the hypothesized link between the true self and free will is indicated by the dashed arrow).

As noted, however, the true self is further aligned with one's essence (see Fig. 1). And it is here—in the possibility that the "true self" is linked to both one's free will and essence—where the mind-body tension arises in full force.

Indeed, people typically consider acts that are committed freely as ones that are not predetermined by material biological causes (e.g., Greene & Cohen, 2004; Nichols, 2011). For example, people are less likely to hold defendants responsible for moral transgressions that are attributable to biochemical and neural factors (Aspinwall, Brown, & Tabery, 2012; Gurley & Marcus, 2008; Heath, Stone, Darley, & Grannemann, 2003; Monterosso, Royzman, & Schwartz, 2005). While people can reconcile physical causes with free will under certain conditions (e.g., Clark, Winegard, & Baumeister, 2019; Nahmias, Morris, Nadelhoffer, & Turner, 2005; Nichols & Knobe, 2007; Roskies & Nichols, 2008), by default, free will is perceived as incompatible with physical causes (Greene & Cohen, 2004). These results suggest that people are intuitive dualists (Bloom, 2004) and that they identify the moral core of agents with the immaterial mind, rather than the material body (see Fig. 1).

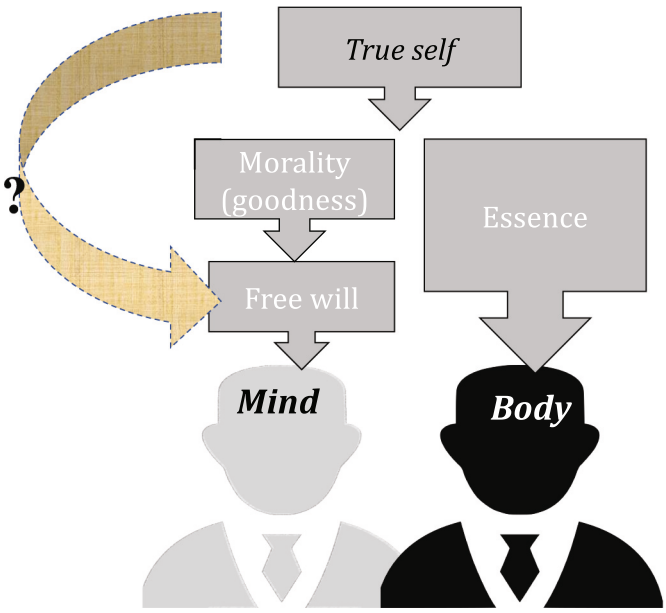

Fig. 1. The true moral me—mind or body?

But per (biological) essentialism, one's essence is not only innately predetermined and immutable (e.g., Keil, 1986), but it must further be anchored in the material body (see Fig. 1). Infants (Setoh, Wu, Baillargeon, & Gelman, 2013) and older children (Gelman, 2003; Gelman & Wellman, 1991) believe that living things must have "insides", that their essence corresponds to a piece of matter (Springer & Keil, 1991) that is localized at the center of the body (Newman & Keil, 2008), and it is linked to some specific biological substance (Waxman, Medin, & Ross, 2007). Moreover, when adults are invited to reason about the origins of a person's psychological traits, they consider traits that are embodied—either in the brain, in the face, or in the internal body—as ones that are more likely to be innate (Berent, Barrett, & Platt, 2020; Berent, Platt, & Sandoboe, in press).

This embodied view of one's (biological) essence—the source of one's "true self", then, would seem to stand in conflict with the notion of the true self as one's good moral core—the putative disembodied home of free will. Fig. 1 captures this tension.

The following research thus asks how we perceive the true self. First is it good or bad, material or immaterial? Second, does the true self correspond to one's essence? Finally, we sought to determine whether judgments of one's moral essence converge with judgments of one's free will.

1.3. The present research

To address these questions, the following experiments presented participants with John – a modern-day reincarnation of Dr. Jekyll and Mr. Hyde. At times, John can be quite positive (e.g., help an elderly person cross the street), but at other times, his acts are negative (e.g., exhibit cruelty to animals).

Participants were told that, on the advice of his family, John

underwent psychological assessment, informed by two tests—a brain test and a behavioral test, and the results of the two tests diverged. In one scenario, the brain test indicated that John's true character was positive whereas the behavioral test suggested it was negative; a second scenario (assigned to another group of participants) supported the opposite conclusion (negative characteristics suggested by the brain test; positive by the behavioral test).

Participants were invited to evaluate John using two different questions. To evaluate John's essence, one question asked participants to determine who John really is, at his core. To gauge John's free will, a second question asked people to consider whether John had committed his positive and negative acts freely. For each question, people rated John's positive and negative attributes separately, on a 1–7 scale.

In so doing, we sought to determine (a) the valence of John's true self—good or bad; (b) whether it corresponds to his mind or body (embodiment); (c) whether it is identified as his essence, and (d) whether it is also aligned with his free will.

The valence of the "true self" is determined by the contrast between John's positive and negative attributes; a "good true self" would be evident if people consistently rated John's positive attributes above his negative ones.

The embodiment of the true self, in turn, was evaluated by the contrast between the brain and behavioral tests. We note that the two tests were strictly matched—all they suggested was whether or not John's response was typical; the brain test offered no additional information (e.g., concerning localization). In the eyes of a dualist, however, the two tests might differ, inasmuch as the brain test explicitly gauges John's body, whereas the behavioral test does not, so the behavioral test could be seen as reflecting John's mind. The contrast between the brain test, then, allowed us to determine whether John's true self (as perceived by participants) is materially embodied or ephemeral: if the true self is perceived as embodied, then people should rate the attributes diagnosed by the brain test (which explicitly references the body) higher than the ones diagnosed behaviorally (a test that could reference the mind); if the true self is perceived as immaterial, then people should rate the outcomes of the behavioral test higher. Finally, the contrast between the free will and essence questions examines whether people align John's true self only with his essence or also with his free will.

1.4. Predictions and implications

In light of the conflicting attributes of John's true self (positive and negative), responses to the "free will" and "essence" questions are open to numerous conflicting predictions. In what follows, we list these possibilities; we flag out the ones we consider most likely.

1.4.1. Essence

In the existing literature, the notion of a good moral core and essence have been considered as intertwined (e.g., De Freitas, Cikara, et al., 2017; De Freitas, Tobia, et al., 2017; Heiphetz, 2019; Strohminger et al., 2017), suggesting that they should both combine to define John's true self. Specifically, if participants indeed perceive John's true self as defined by his essence, and if they further perceive the true self as morally good, then, in this characterization of the true self, participants should consider John's good acts as more representative of his essence. If participants also perceive John's essence as embodied, then acts that are demonstrably embodied (i.e., detectable by a brain test) should be further considered as more representative of his essence. Moreover, the two factors—goodness and embodiment—could combine synergistically to render good acts that are detected by the brain test as the ones that are most representative of John's essence.

But if the two attributes of one's moral essence—its goodness and embodiment—stand in mutual tension, then when the two factors are put in conflict, they may not combine synergistically, but rather bifurcate. Indeed, our experiments systematically pit goodness and embodiment against each other (e.g., the brain test suggests either good or bad attributes, not both), and past research has shown that, when people consider biological essence, materiality is paramount (Berent, Barrett, & Platt, 2020; Berent, Platt, & Sandoboe, in press). We thus hypothesize that, when people evaluate one's moral essence, and goodness and embodiment conflict, embodiment trumps. If so, then participants should rate acts that are demonstrably embodied (i.e., those that manifest in the brain) as more indicative of John's essence, and this should be the case irrespective of valence (good or bad).

1.4.2. Free will

Responses to the "free will" question are likewise open to multiple interpretations. One possibility is that people will simply align their judgments of John's free will with his perceived essence. If so, responses to the "free will" question will converge with responses to the "essence" question (as discussed above), and since past research has argued that one's essence speaks to one's true self, the notions essence and free will should converge, and they should both align with John's true self.

In other scenarios, people will evaluate each question independently, and if so, responses to the two questions need not converge. In fact, unlike the essence question, free will responses may not even speak to the true self at all. Indeed, past research has shown that people invoke free will to explain both positive (e.g., Baumeister, Masicampo, & DeWall, 2009; Seto & Hicks, 2016) and negative acts (e.g., Baumeister et al., 2009; Martin, Rigoni, & Vohs, 2017; Shariff et al., 2014; Vohs & Schooler, 2008). These results make it clear that people know too well that the self is the source of both good and bad acts, and that these acts are committed freely. Accordingly, when asked to evaluate John's conflicting acts, it is not clear, a priori, whether participants will attribute them to John's self (the source of all acts, good or bad) or to his true self (presumably, the source of his good acts only).

If participants align free will with John's self (as opposed to his true self), then they should be equally likely to consider good- and bad acts as freely willed. And since free will is typically considered unbound by bodily causes (e.g., Greene & Cohen, 2004; Nichols, 2011), the most likely scenario, here, is that people should align John's free will with the outcomes of behavioral test (as this test does not explicitly reference the body), and, as noted, this should be the case irrespective of valence (for both good and bad acts).

On a third scenario, however, responses to the free will question could potentially speak not to John's self, generally, but to his true self specifically. Indeed, participants might be intrigued by John's conflicting moral attitudes, and past research has shown that, when a character presents conflicting moral attributes, people seek to determine the protagonist's true self (e.g., Strohminger et al., 2017). By interrogating which of John's acts is freely willed, participants might thus seek to unveil his presumed underlying moral core—his true self. Given that the true self is defined morally, and that morality is linked to free will, we consider this last scenario as the most likely.

If people do indeed align John's free will with his true self (as opposed to the self), and if they further consider the true self as good, then they should consider John's good acts as more likely to be committed freely than his bad acts. Considering the effect of test, recall that free will judgments typically do not reference the body, but when free will and bodily causes are in conflict (as in the case here), people are willing to uphold free will even when bodily causes are present (Clark et al., 2019; Nahmias, Shepard, & Reuter, 2014). We thus expect that, when the moral valence of the true self and its embodiment are in conflict, valence (specifically, goodness) trumps.

We note that this expectation is in full parallelism to our predictions for the essence question above. Each such question is informed by two attributes (goodness and embodiment), and when these attributes are in conflict, they bifurcate, such that the attribute most relevant to the probe in question wins. But while, for the essence of the true self, the putative "winner" is the body, for the moral true self, it is valence, specifically, goodness, that has the upper hand. We thus hypothesize that, seen as one's (biological) essence, the true self is necessarily

embodied (goodness is preferred, but not obligatory); seen as one's moral core (specifically, as the home of one's free will), the true self is necessarily good (here, disembodiment is preferred, but not obligatory).

Computationally, this hypothesis suggests that, each component of the true self (essence and morality) is evaluated in terms of two violable constraints—goodness vs. (dis)embodiment; the two components, however, differ with respect to the ranking of the two constraints. The computation of essence outranks embodiment over goodness; the computation of morality via free will exhibits the opposite ranking. We capture this computational hypothesis in (1), inspired by the framework of Optimality Theory (Prince and Smolensky, 1993/2004).

(1) The hypothesized ranking of goodness and embodiment constraints in the evaluation of essence and free will
   a. Essence: embodiment > goodness
   b. Morality (free will): goodness > (dis)embodiment

Altogether, then, we predict that, in the free will question, people should rate John's positive attributes higher than his negative attributes (irrespective of test), whereas in the "essence" question, they should rate the attributes diagnosed by the brain test higher than the ones diagnosed behaviorally (irrespective of the act's valence).

Experiments 1a-c explore the effect of test (brain/behavior) and act valence (good/bad)—these three studies differ only on the precise wording of positive and negative acts. Experiments 2–3 next examine whether the divergence in response to the free will and essence questions depends on the embodiment of John's characteristics (as suggested by the test results).

Our investigation sheds light on the intuitive psychological notion of the "true self" (e.g., De Freitas, Cikara, et al., 2017; Newman et al., 2014a; Strohminger et al., 2017); on laypeople's moral judgments (e.g., Barrett et al., 2016; Greene & Cohen, 2004; Nichols, 2011) and their anchoring in intuitive essentialism (e.g., De Freitas & Cikara, 2018; De Freitas, Cikara, et al., 2017; Heiphetz, 2019; Newman et al., 2014b; Strohminger et al., 2017; Strohminger & Nichols, 2014) and dualism (Bloom, 2004); and on the role of neuroscience in informing these judgments (e.g., Hopkins, Weisberg, & Taylor, 2016; Weisberg, Hopkins, & Taylor, 2018; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008; Weisberg, Taylor, & Hopkins, 2015).

2. Experiment 1a (Benevolence vs. Aggression)

2.1. Methods

2.1.1. Participants

Eighty participants were assigned to Experiment 1a; one participant in Experiment 1 did not complete the experiment, resulting in a total of 79 participants.

In Experiments 1a-c, participants were recruited from Amazon Mechanical Turk; participants in Experiments 2–3 were sampled from Prolific. Participants were all adult native English speakers who were reportedly free of language and reading disorders.

In Experiment 1a, participants had also reportedly not taken any advanced courses in psychology (100%), linguistics (100%), and many had not taken advanced courses in biology (82%). Participants reported their highest levels of education as follows: 29% high school, 58% college, 11% a graduate school program, and 1% none (beyond elementary-middle school).

Sample size in this and all subsequent experiments was determined a priori (before any data analysis) by a power sensitivity analysis based on pilot results. A sensitivity power analysis (a *t*-test of a linear regression model, single coefficient) suggested that the chosen sample size (N=80) has 0.80 probability to detect a correlation of 0.30 among the model's three predictors at the alpha level of 0.05 (i.e., f2=0.10). Experiments 1 (a-c)-3 each included its own distinct group of participants.

2.1.2. Materials and procedure

Participants in Experiment 1a read one of two vignettes, each describing John—a character whose behavior is inexplicably erratic, altering between acts of benevolence (e.g., stopping to help an elderly person cross the street and donating money to the needy) and aggression (e.g., cruelty to animals and randomly bullying students on his College campus). At the advice of his family, John approaches a psychologist who evaluates his condition using both behavioral and brain tests. The results of the two tests are in conflict, and the nature of the conflict varied for the two vignettes. In one vignette (assigned to half of the participants), the brain results indicated that John was benevolent and the behavioral test provided evidence for aggression. A second vignette (assigned to the second half of participants) presented the opposite scenario (the behavioral test indicated benevolence, whereas the brain test indicated aggression).

Participants were next asked to address two questions (with order counterbalanced). One question invited people to reason whether "When John commits acts of benevolence/aggression, he performs those acts of his own free will". Another question asked people reason about John's "real essence". Specifically, they were to determine whether, "at his core, John is a benevolent/aggressive person". Participants provided their responses on a 1–7 scale (1=strongly disagree; 2=disagree; 3=slightly disagree; 4=neither agree nor disagree; 5=slightly agree; 6=agree; 7=strongly agree). For each such question, people responded to the "benevolence" and "aggression" probes separately. All measures, manipulations, and exclusions in the study are reported.

2.2. Results and discussion

Fig. 2 presents participants' responses to the "free will" and "essence" questions. In this and all subsequent experiments, we plot the results by act (benevolence vs. aggression) and by the test whose outcome was congruent with that act. For example, the benevolence/brain bar indicates responses to the benevolence probe given a brain test suggesting that John is benevolent (and a behavioral test suggesting he is aggressive). By the same token, the benevolence/behavior bar reflects responses to the benevolence probe given a behavioral test suggesting that John is benevolent (and a brain test suggesting he is aggressive).

An inspection of the means suggests that responses to the free will and essence questions diverged. When asked to reason about John's free will (Fig. 2A), people based their responses on the act—they considered benevolent acts as more likely to be freely committed compared to acts of aggression. But when people evaluated John's essence (Fig. 2B), they then based their responses on the test, and they considered the brain test as more indicative of John's essence than the behavioral test.

These conclusions were supported by a linear mixed effects model with random intercept by participants (lmer (rating ~ Test * Act+ (1| Participant)) using sum coding. To clarify—the Test factor indicates which test (brain or behavior) is congruent with a given act. For the Test factor, the sum coding was Brain= −0.5, Behavior = 0.5; for the Act factor, it was Benevolent = −0.5, Aggressive = 0.5.[1]

In what follows, we report regression output in terms of the beta estimates (β), standard errors, t statistics, and *p*-values as reported by the lmer() function in R, in addition to reporting standardized beta coefficients (β) and their 95% confidence intervals (CI). Standardized beta coefficients and CIs were calculated using the standardized_parameters function of the effectsize package (Ben-Shachar, Makowski, & Lüdecke, 2020) in R.

Results for the "free will" question yielded only an effect of Act

[1] Because, in this design, each participant contributed only two observations per cell, some of these models failed to calculate random effects, and, by default, fell back to a linear regression model (without random effects). An analysis of the results using linear regression (without random effects) supported the same conclusions.

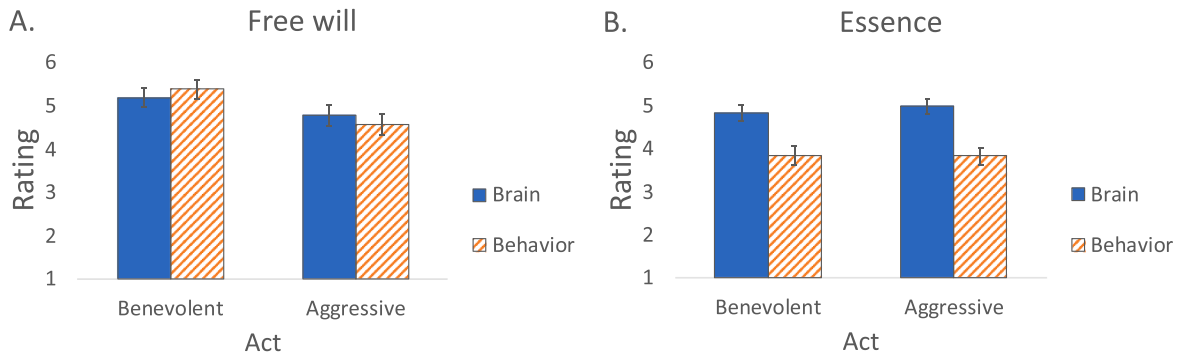

Fig. 2. Responses to the free will (A) and essence (B) questions in Experiment 1a. Error bars in this and all figures are standard error.

(*β=−0.61, SE=0.21, t(77)=−2.89, p*=.01; β = −0.41, CI = −0.69 – -0.13). The effect of Test (*β=−0.01, SE=0.21 t(77)=−0.04, p=.97;* β = −0.0001, CI = −0.28–0.27) and the Test x Act interaction (*β=−0.41, SE=0.51, t(77)=−0.80, p=.43;* β = −0.27, CI = −0.94–0.40) were not significant.

For the "essence" question, we found only a reliable effect of Test (*β=−1.08, SE=0.20, t(154)=−5.36, p<.001;* ß = −0.79, CI = −1.08 – -0.50). The effects of Act (*β=0.08, SE=0.20, t(154)=0.37, p=.71;* ß = 0.06, CI = −0.23–0.34) and the interaction (*β=−0.16, SE=0.40, t (154)=−0.40, p=.69;* ß = −0.12, CI = −0.69–0.46) were not significant.

These results demonstrate for the first time that the evaluation of a person's moral behavior dissociates, depending on the question (free will vs. essence) and the diagnostic test (brain vs. behavior).

To ensure that these conclusions are not due to the strong emotive connotations of aggressive acts, Experiments 1b-c replicate the findings using two other contrasts (benevolence/malice; kindness/unkindness). To further counter the possibility that people might have interpreted John's volatile personality as abnormal (hence, unrepresentative of typical individuals), we invited participants to explain their ratings. If people indeed hold conflicting notions of John's moral character, then similar results should obtain when John's actions are devoid of emotive and abnormal connotations.

3. Experiment 1b (Benevolence vs. Malice)

3.1. Methods

3.1.1. Participants

Eighty participants took part in Experiment 1b. Participants had reportedly not taken any advanced courses in psychology (100%), linguistics (100%), and many had not taken advanced courses in biology (85%). Participants reported their highest levels of education as follows: 41% high school, 43% college, 16% a graduate school program, and 0% none (beyond elementary-middle school).

3.1.2. Materials and procedure

The materials and procedure were as in Experiment 1a, except that the positive and negative acts were described as benevolence vs. malice. Additionally, at the end of the experiment, participants were prompted to provide a brief explanation for their responses (for the materials, see Appendix I). All measures, manipulations, and exclusions in the study are reported.

3.2. Results and discussion

As in Experiment 1a, people were more likely to identify John's free will with his good acts, but they aligned his essence with the outcome of the brain test (see Fig. 3).

For the free will question, the regression yeilded only a significant effect of Act (*β=−0.61, SE=0.17, t(78)=−3.60, p*<.0006; ß = −0.41, CI = −0.63 – -0.19). The effects of Test (*β=0.24, SE=0.17, t(78)=1.40, p*=.17; ß = 0.16, CI = −0.06–0.38) and the interactions (*β=0.23, SE=0.57, t(78)=0.40, p*=.69; ß = 0.15, CI = −0.59–0.89) were not signficant.

The essence question, by contrast, only yielded an effect of Test (*β=−0.60, SE=0.22, t(156)=−2.67, p*=.01; ß = −0.41, CI = −0.72 – -0.11). The effects of Act (*β=−0.35, SE=0.22, t(156)=−1.56, p*=.12; ß = −0.24, CI = −0.54–0.06) and the interaction (*β=0.20, SE=0.45, t (156)=0.45, p*=.66; ß = 0.14, CI = −0.47–0.74) were not signficiant. These conclusions were unchanged when we excluded from the analysis the minority of pariticipants (*N*=13) whose written responses implied that John suffered from psychologcial abnormlaities (see, SM).

4. Experiment 1c (Kindness vs. Unkindness)

4.1. Methods

4.1.1. Participants

Eighty participants took part in Experiment 1c. Participants had

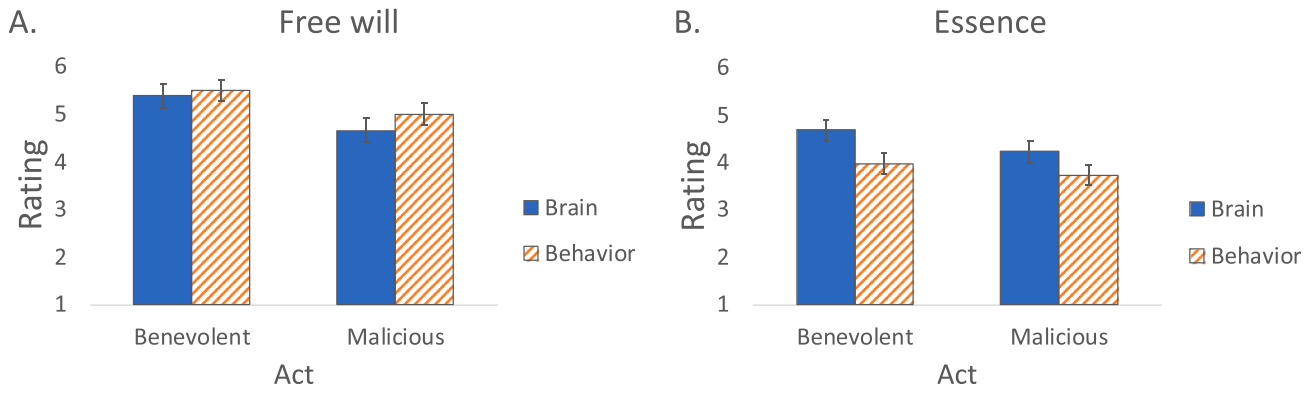

Fig. 3. Responses to the free will (A) and essence (B) questions in Experiment 1b. Error bars in this and all figures are standard error.

reportedly not taken any advanced courses in psychology (100%), linguistics (100%), and many had not taken advanced courses in biology (86%). Participants reported their highest levels of education as follows: 43% high school, 45% college, 13% a graduate school program, and 1% none (beyond elementary-middle school).

4.1.2. Materials and procedure

The materials and procedure were as in Experiment 1b, except that the positive and negative acts were described as kindness vs. unkindness (for the materials, see Appendix I). All measures, manipulations, and exclusions in the study are reported.

4.2. Results and discussion

An analysis of participants' written responses to the kind/unkind contrast identified a substantial group (N=26) who ascribed to John psychological abnormality. Thus, we first provide the results for all participants; we next reanalyze the findings after excluding participants who referenced psychological abnormalities.

All participants. As in previous experiments, people were more likely to identify freely-willed acts as positive (see Fig. 4). The analysis of the free will question indeed yielded only a significant effect of Act (ß=−0.90, SE=0.21, t(78)=−4.21, p<.001; ß = −0.58, CI = −0.85 – -0.31). The effects of Test (ß=−0.08, SE=0.21, t(78)=−0.32, p=.73; ß = −0.05, CI = −0.32–0.22) and the interaction (ß=−0.35, SE=0.51, t (78)=−0.69, p=.50; ß = −0.23, CI = −0.87–0.42) were not significant.

Also in line with previous experiments, people identified John's essence with the outcome of the brain test. The effect of Test (*ß=−0.73, SE=0.22, t(156)=−3.24, p*<.002; ß = −0.49, CI = −0.78 - −0.19) was highly significant, and it was not further modulated by Act (*ß=−0.10, SE=0.45, t(156)=−0.22, p=.82,* ß = −0.07, CI = −0.65–0.52). But unlike previous experiments, here, we also found a significant effect of Act (*ß=−0.73, SE=0.22, t(156)=−3.24, p*<.002; ß = −0.49, CI = −0.78 - −0.19), as good acts were considered more likely to indicate of John's essence. This outcome could have arose because some participants viewed John as mentally ill, so they might have been reluctant to attribute his unkind acts to his essence. To counter this possibility, we next repeated the analysis for participants whose responses were free of any implied disorder.

Disorder free responses. The "disorder free" results (Fig. 5) fully converged with previous experiments. Freely willed acts were more likely to be identified as good, whereas those that define John's essence were aligned with the brain test.

For the free will question, the effect of Act was significant (ß=−0.52, SE=0.25, t(52)=−9.09, p=.04; ß = −0.35, CI = −0.68–0.02). The effects of Test (ß=−0.18, SE=0.25, t(52)=−0.74, p=.46; ß = −0.12, CI = −0.45–0.20) and the interactions (ß=−0.17, SE=0.63, t(52)=−0.26, p=.79; ß = −0.11, CI = −0.95–0.73) were not significant. The essence question, by contrast, now yielded a significant effect of Test (ß=−1.23, SE=0.26, t(104)=−4.71, p<.001; ß = −0.84, CI = −1.19 – -0.49). The effects of Act (ß=0.03, SE=0.26, t(104)=0.10, p=.92; ß = 0.02, CI = −0.33–0.37) and the interaction (ß=−0.40, SE=0.52, t(104)=−0.77, p=.44; ß = −0.27, CI = −0.97–0.42) did not approach significance.

These results confirm that the dissociation between the perception of a person's free will and their essence is a robust phenomenon that is independent of the specific wording of positive and negative acts. Free will is linked to a person's good acts, whereas their essence is aligned with the outcome of the brain test.

We suggest that people aligned John's essence with the brain test outcome because, per essentialism, one's essence must be materially embodied, and for the Dualist, only the brain test offers explicit evidence for embodiment. Experiments 2–3 further examine this hypothesis.

5. Experiment 2 (no test results)

Experiment 1a-c showed that, when our protagonist's moral characteristics conflict, people evaluate his moral core differently, depending on whether the conclusions are based on a brain- or a behavioral test.

We hypothesize that the divergence arises because the true self references two conflicting notions (morality and essence), and each such notion, in turn, differs, with respect to how it ranks the goodness of true self and its embodiment (restated in (1), below). Seen as one's moral core, the true self must be good (disembodiment is not obligatory); seen as one's (biological) essence, it must be embodied (goodness is not obligatory). Since the test results offer evidence for the embodiment of John's characteristics, and since it further pits their embodiment against their valence, it is the test that forces participants to choose between these conflicting attributes, and in so doing, it promotes the divergence between John's essence and free will.

(1) The hypothesized ranking of goodness and embodiment constraints in the evaluation of essence and free will

a. Essence: embodiment > goodness
b. Morality (free will): goodness >(dis)embodiment

If this analysis is on the right track, then once the test results are removed, the divergence between the free will and essence probes ought to be eliminated, and people should now rate John's good acts higher than negative acts for both the essence and free will questions. In contrast, once the test results are reintroduced, the divergence between free will and essence should re-emerge, even when participants are presented with no further information about John's character. The first manipulation (the removal of the test), then, should show that the test results are *necessary* to elicit the divergence between the free will and essence questions; the second manipulation (the test results alone) should show that the test results are *sufficient*. These questions are addressed in Experiments 2–3, respectively.

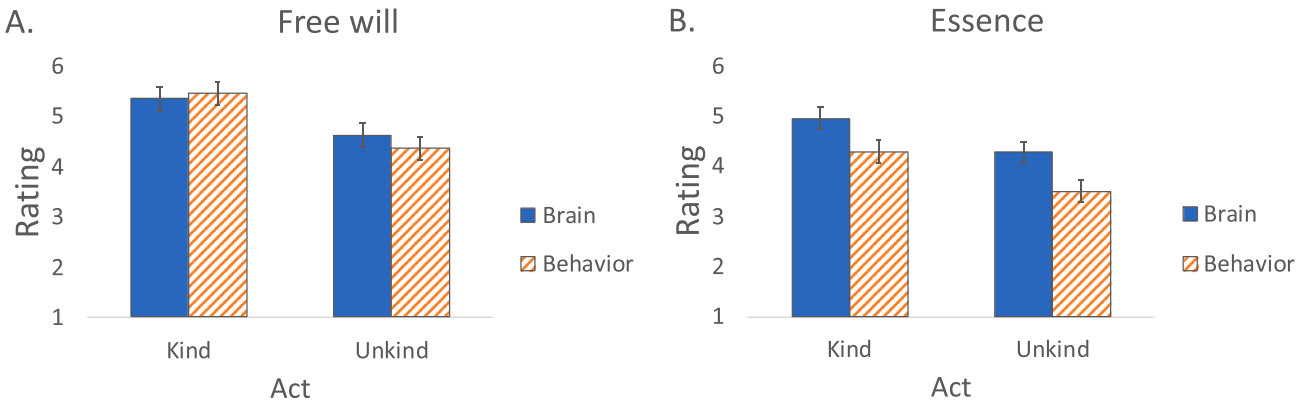

Fig. 4. Responses to the free will (A) and essence (B) questions from the entire sample in Experiment 1c. Error bars in this and all figures are standard error.

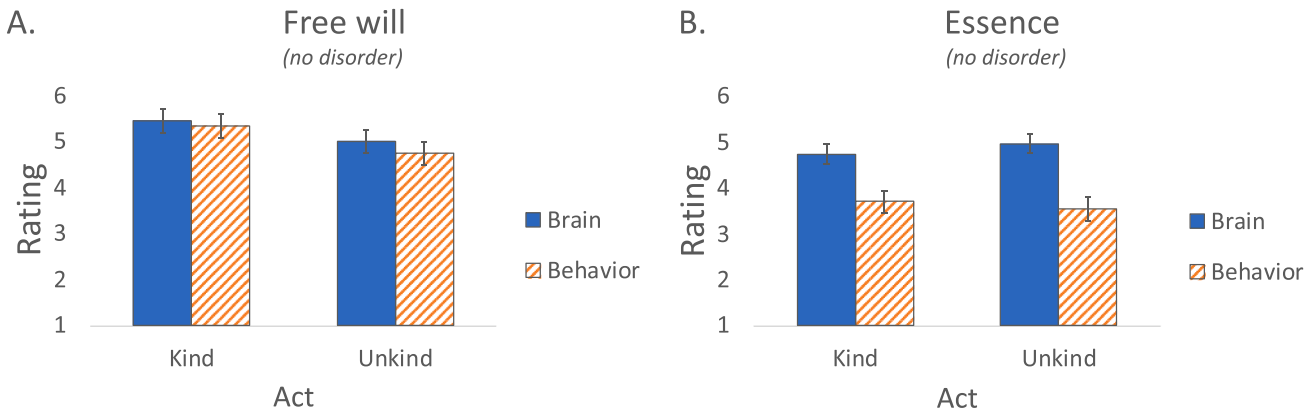

Fig. 5. Responses to the free will (A) and essence (B) questions from participants who did not ascribe to John psychological abnormalities in Experiment 1c. Error bars in this and all figures are standard error.

Experiment 2 removed the test outcomes. Participants, here, were only presented with a minimal description of John's conflicting personality characteristics (at times; John is kind; at other times he can be aggressive); a separate control condition (reported in the SM) confirmed that, when this description is paired with the test outcomes (as in Experiment 1), responses to the free will and essence questions indeed diverge, as expected. The critical question, then, is whether the removal of the test outcomes in Experiment 2 will eliminate the divergence between the two questions.

If the conflict we had observed between the free will and essence questions arises from the embodiment of John's characteristics (as indicated by the conflict between brain and behavioral tests), then once the test results are removed, the divergence between the free will and essence questions should be eliminated. Assuming further that responses to the two questions reference John's good true self, we now expect that responses to the free will and essence questions should converge, and they should both side with John's positive behaviors (i.e., with his good true self).

Experiment 3, next, evaluated whether the test outcomes are sufficient to elicit the different responses. Here, we reintroduced the outcomes of the brain and behavioral tests, but offered no other information about John's condition (i.e., the opening description of John's conflicting personality characteristics and acts was removed). If the divergent responses to the free will and essence questions indeed result from the conflicting test outcomes (i.e., the test outcomes are sufficient to produce the divergence), then once the outcomes of the two tests are given, responses to the two questions should once again diverge.

5.1. Methods

5.1.1. Participants

Eighty participants took part in Experiment 2. Many participants had reportedly not taken any advanced courses in psychology (69%), linguistics (78%) and biology (61%). Participants reported their highest levels of education as follows: 43% high school, 46% college, 11% a graduate school program, and 0% none (beyond elementary-middle school). Of this sample, 51% of participants were female, 48% were male, and 1% preferred not to disclose gender; the mean age was 24.3 years (SD=5.63).

5.1.2. Materials and procedure

The materials were as in Experiment 1a, except that the shifts in John's character were now described in terms of his conflicting personality attributes, rather than his specific acts (for the materials, see Appendix I). Critically, the test information was eliminated. Participants were asked to help the psychologist evaluate John's free will and essence (for the materials, see Appendix I). All measures, manipulations, and exclusions in the study are reported.

5.2. Results and discussion

Experiment 2 examined whether the test outcomes are necessary to elicit the divergence between the free will and essence questions. If they are, then once the test results are eliminated, the divergence should disappear. If by default, participants believe that John's true self is good, and if the true self is referenced by the essence question, then participants should now uniformly side with John's good acts, as they do in assessing John's free will. Fig. 6 provides the mean response to the "free will" and "essence" questions (please note that, here, not test is provided).

Results are consistent with this prediction. The regression results (lmer(rating ~ Act + (1|Participant)) yielded a reliable effect of Act for both the free will (β=−0.78, SE=0.16, t(79)=−5.00, p<.0001; β = −0.59, CI = −0.82 - −0.36) and essence (β=−0.85, SE=0.17, t(158)=−4.93, p<.0001; ß = −0.73, CI = −1.02 - −0.44) questions.

These results suggest that the conflicting test outcomes were necessary to elicit the divergence between the questions, as once these outcomes were removed, people considered John's good acts not only as more indicative of his free will but also of his essence—in line with previous research on the moral true self (Heiphetz et al., 2017).

To determine whether the test outcomes are further sufficient to elicit the divergence, Experiment 3 once again compares responses to the free will and essence questions when all information about John's background is removed. In so doing, we further sought to counter an alternative explanation for the results of Experiments 1a-c. In this view, the divergence between the free will and essence questions emerged simply because these experiments informed participants of John's conflicting behaviors, and this information led participants to partly distrust the behavioral test. This alternative explanation predicts no dissociation between the two tests in Experiment 3 (once participants are provided with no information about John's conflicting behaviors). But if, contrary to this suggestion, the divergence was indeed caused by the test outcomes (specifically, their perceived embodiment), then the test results should be sufficient to elicit the same pattern even when all other information about John is eliminated (in Experiment 3).

6. Experiment 3 (only test results)

6.1. Methods

6.1.1. Participants

Two hundred participants took part in Experiment 3. Sample size, here, was increased to ensure the minimum of 50 participants per cell (Simmons, Nelson, & Simonsohn, 2018). Many participants had

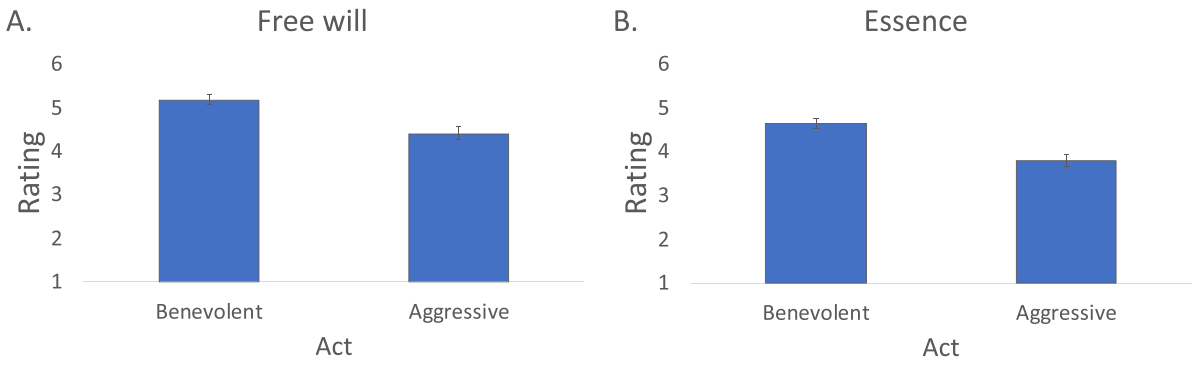

Fig. 6. Responses to the free will (A) and essence (B) questions in Experiment 2. Error bars in this and all figures are standard error.

reportedly not taken any advanced courses in psychology (61%), linguistics (85%) and biology (56%). Participants reported their highest levels of education as follows: 32% high school, 57% college, 12% a graduate school program, and 0% none (beyond elementary-middle school). In this sample, 49.5% of participants were female, 50.5% were male; the mean age was 24.5 years (SD=6.18).

6.1.2. *Materials and procedure*

Materials and procedure were identical to those in Experiment 1a, except that here, the vignette entirely eliminated the discussion of John's condition. To motivate the testing and results, participants were simply informed that "John exhibits an erratic personality" (for the materials, see Appendix I). All measures, manipulations, and exclusions in the study are reported.

6.2. *Results and discussion*

An inspection of the results (Fig. 7) suggests that, once the test results were introduced, the divergence between responses to the free will and essence questions remerged, as in Experiments 1a-c.

The analysis of the free will question indeed yielded only a significant effect of Act (β=−0.37, SE=0.14, t(198)=−2.55, p=.01; ß = −0.24, CI = −0.43 – -0.06). The effects of Test (β=−0.11, SE=0.14, t(198)=−0.73, p=.46; ß = −0.07, CI = −0.26–0.12) and the interaction (β=0.03, SE=0.31, t(198)=0.10, p=.92; ß = 0.02, CI = −0.38–0.42) were not significant. In contrast, for the essence question, we only found a significant effect of Test (β=−1.09, SE=0.13, t(396)=−8.27, p<.0001; ß = −0.77, CI = −0.95 – -0.58). The effect of Act (β=−0.16, SE=0.13, t(396)=1.18, p=.24; ß = 0.11, CI = −0.07–0.29) and the interaction were not significant (β=−0.03, SE=0.26, t(396)=−0.11, p=.91; ß = −0.02, CI = −0.38–0.34).

Thus, for the essence question, participants sided with the brain test (i.e., they rated the outcomes diagnosed by the brain test higher than the ones diagnosed behaviorally); for the free will question, they favored John's good acts (i.e., they considered John's good acts as more likely to be committed freely than bad ones). The re-emergence of this pattern despite no additional information on John's condition demonstrates that the test outcomes (specifically, the information they provide with respect to the embodiment of John's characteristics) are not only necessary but also sufficient to elicit these divergent findings. These results further show that the preference for the divergent outcomes of the brain and behavioral tests is inexplicable by the description of John's previous acts (i.e., the possibility that participants in Experiments 1a-c disregarded the behavioral test because they were informed that John's behavior is erratic). Together, these findings demonstrate that the evaluation of a person diverges, depending on (a) whether participants consider one's free will or essence; and on (b) whether the act is diagnosed by a brain test or behaviorally.

7. General discussion

Experiments 1–3 examined whether people hold a notion of the true self, distinct from the self, and evaluated its characteristics. We asked whether people identify a person's true self with one's biological essence, and whether they align it with one's mind or body. We also explored how the notion of one's essence relates to the perceptions of one's free will.

To this end, we invited participants to reason about John's conflicting acts—positive and negative, based on two tests with conflicting results – a brain test and a behavioral test (reflecting John's body and mind, respectively). We asked participants to evaluate John's essence, and to decide which of his acts were committed freely.

Results showed that responses to these questions diverged. When participants judged John's essence, they consistently sided with the outcomes of the brain- over the behavioral test. Moreover, participants invariably rated the act congruent with the brain test outcome above the scale's "neutral" midpoint, whereas this was not the case for the outcome of the behavioral test (see SM, Table S1). Thus, not only was the brain test outcome considered more representative of John's essence, but when judged in absolute terms, people aligned John's essence only

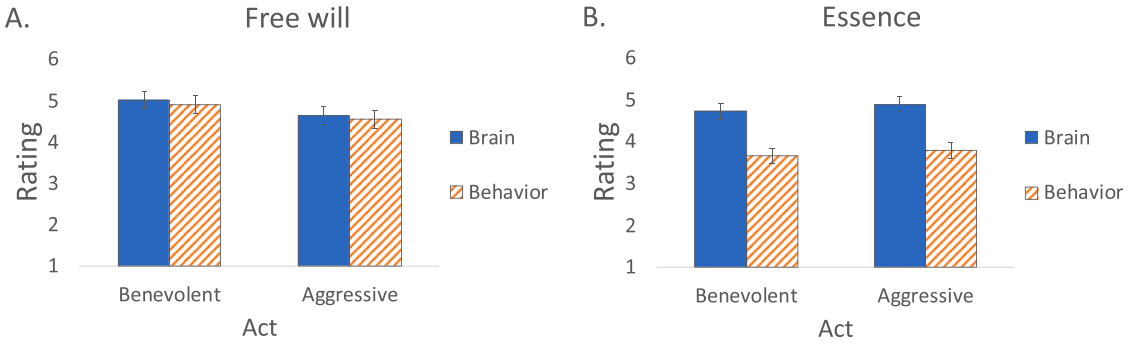

Fig. 7. Responses to the free will (A) and essence (B) questions in Experiment 3. Error bars in this and all figures are standard error.

with the brain, not the behavioral test. With a single exception (due to a minority of participants in Experiment 1c who perceived John as mentally ill), "essence" judgments were unaffected by the act's valence (good or bad). Thus, participants typically aligned John's essence with his brain.

But when people evaluated John's free will, here, they consistently favored "good" over "bad" acts. This is not necessarily because participants outright negated that John's bad acts were freely willed. Indeed, participants rated all acts—good or bad—higher than the scale's midpoint (see SM, Table S1). This is in line with previous research, suggesting that participants view both positive and negative acts as freely willed (Baumeister et al., 2009; Martin et al., 2017; Shariff et al., 2014; Vohs & Schooler, 2008). Also in line with past research, participants in our experiments further considered John's good acts as freely willed regardless of test outcomes, even when they were linked to his brain (i.e., body Clark et al., 2019; Nahmias et al., 2014). But when good acts were contrasted with bad ones, participants considered John's good acts as more likely to be freely willed. Moreover, unlike the perceptions of John's essence, the evaluation of his free will was independent of test.

The discrepancy is puzzling, given that in existing accounts, the two characteristics of "true self"— as a (good) moral core and the home of one's (biological) essence—are seen as seamlessly intertwined (e.g., De Freitas & Cikara, 2018; De Freitas, Cikara, et al., 2017; Heiphetz, 2019; Newman et al., 2014b; Strohminger et al., 2017; Strohminger & Nichols, 2014). Since moral appraisal typically references free will (e.g., Nichols, 2011; Nichols & Knobe, 2007; Roskies & Nichols, 2008; Sarkissian et al., 2010) and intentionally (Barrett et al., 2016; Greene & Cohen, 2004; Nichols, 2011), one would have just expected John's free will and essence to interact synergistically. Accordingly, participants should have gauged John's free will by simply back-computing from his essence, and thus, arrive at similar responses to the two questions. This, however, is not what was found.

It is unlikely that these conclusions obtained because people perceived John as suffering from a clinical abnormality. First, the abnormality presumption fails to explain why the evaluation of John's true self depends on the outcome of the two Tests—brain and behavioral, and the Question—free will vs. essence. Second, the pattern above was obtained even when we removed the results of participants who implied that John suffered from psychological abnormality.

It is also unlikely that the results are due to the specific characterization of good and bad behaviors (in Experiments 1a-c) or to the abrupt changes in his behavior. Experiment 3 makes it clear that the same results obtain when the description of John's specific acts and personality characteristics is eliminated entirely.

Why then, did responses to the two questions (free will and essence) diverge? And why did responses to the essence (but not free will) question depend on the type of test—brain and behavior?

We suggest that these conflicting results reflect different competing facets of a common construct—the "true self". The possibility that responses to the "essence" questions reflect John's true self is in line with the past findings, suggesting that the true self arises from essentialist reasoning (De Freitas & Cikara, 2018; De Freitas, Cikara, et al., 2017; Heiphetz, 2019; Newman et al., 2014b; Strohminger et al., 2017; Strohminger & Nichols, 2014). Since per biological essentialism, one's essence is aligned with the body (Newman & Keil, 2008; Springer & Keil, 1991; Waxman et al., 2007), the judgments of one's essence—the true self—are thus expected to reference the body. The consistent preference to identify John's essence with the brain test (which explicitly references the body) is in line with this possibility. These results also agree with our past research, showing that, when psychological traits "show up" in a brain test, people are more likely to consider these traits as inborn (hence, as more indicative of one's biological essence) compared to the when the same traits are gauged behaviorally (Berent, Barrett, & Platt, 2020; Berent, Platt, & Sandoboe, in press). Together, these results seem to suggest that, when participants consider one's (biological) essence, they view the true self as materially embodied.

We suggest that, notwithstanding the different outcomes, responses to the "free will" question may well have likewise targeted the true self. As noted in the Introduction, a priori, this is certainly not the only possible prediction. Participants could have based their judgment of free will on the "essence" probe, or alternatively, considered only which of John's acts was committed freely—without referencing John's true self. But our results are inconsistent with these scenarios. Had participants sided with "essence", they should have favored the brain results. Similarly, had they considered the self's free will (rather than the true self, specifically), they should have been equally likely to select both acts, as free will is demonstrably relevant to both positive and negative acts committed by agents (Baumeister et al., 2009; Martin et al., 2017; Shariff et al., 2014; Vohs & Schooler, 2008). Neither of these conclusions is borne out.

We thus submit that free will judgments targeted John's true self. We speculate that participants were intrigued by John's erratic behavior, and they sought to determine John's underlying moral core by referencing John's true self—his good moral core. The appraisal of John's true self, then, led them to be more likely to consider his good acts as freely willed. These findings, then, open up the possibility that one's true self is not only morally good but is further explicitly identified with one's free will.

But if responses to the "free will" and "essence" questions both referenced the same tacit construct—the true self, then why did responses to the two questions diverge?

We suggest that the divergence emerged because these two questions differentially weigh the information presented by the test with respect to the embodiment of John's attributes and their valence. In line with this proposal, responses to the two questions differed only when the test outcomes were distinct (in Experiments 1 & 3), and this was the case even when information about John's character was eliminated (in Experiment 3). But when the test was removed (in Experiment 2), responses to the free will and essence questions converged, as they both preferentially referenced John's better self.

We thus conclude that the two questions—the essence and free will—each reference the same tacit notion of the "true self", but they are informed by distinct computations. Each such computation, in turn, differs on how it ranks the goodness of the true self and its embodiment. The computation of (biological) essence only requires that the true self be materially embodied (goodness is not obligatory); that of free will only requires that the true self be good (its disembodiment is not obligatory); the ranking is restated in (1), below). Thus, when the test results pitted goodness and embodiment against each other, the two constructs bifurcated, thereby revealing their underlying tension.

(1) The hypothesized ranking of goodness and embodiment constraints in the evaluation of essence and free will

a. Essence: embodiment > goodness
a. Morality (free will): goodness >(dis)embodiment

This tension between the "embodied" and "disembodied" true self is also evident in the literature. One set of results suggests that the true self is immaterial. For example, people believe that the self continues to exist even when the person's memories are implanted in a robot (Blok, Newman, Behr, & Rips, 2001). In fact, people extend the notion of the true self even to entities that are devoid of any material instantiation at all—to groups (true nations, universities, and bands; De Freitas, Tobia, et al., 2017) and emotions (e.g., happiness, Phillips, De Freitas, Mott, Gruber, & Knobe, 2017). Other results, however, suggest that the material body does inform reasoning about the true self. For example, people are more likely to conclude that a person's true self transfers to a robot if the transplant includes the person's brain (compared to their memories alone, Blok et al., 2001). Similarly, if a transplant preserves the body, people believe implicitly that the self persists (even if memory is lost, Nichols & Bruno, 2010).

We thus suggest that, when people consider John's free will, people evaluate John's goodness, irrespective of his material body; when they consider his essence, they inspect his body. Consequently, when considering free will, they conclude that John is good, irrespective of whether the evidence for goodness is explicitly linked to the body (by the brain test) or not (behaviorally); but when they consider John's essence, here, it is his material properties that are paramount, and for this reason, the outcomes of the brain test trump. These results suggest that our tacit notion of the true self is not unitary.

To be clear, these conflicting *tacit* notions of the true self need not correspond to people's *explicit* judgments. Indeed, when asked to explicitly reflect on who they are, we would fully expect people to insist that their true self is unitary, and even resist claims to the contrary. Our results, however, suggest that people's tacit notions of the true self might be conflicting. If so, our explicit psychological belief in a true unitary "me" may be illusory.

The contrast we have unveiled between our explicit notion of the true self as unitary, and the distinct, and at times, conflicting computations that inform its tacit evaluation mirrors the well-known dissociation between explicit and implicit evaluation of social attitudes (e.g., Dasgupta, 2004; Kurdi et al., 2019; Phelps et al., 2000). As in the evaluation of social stereotypes, conclusions regarding the true self diverge, depending on how these attitudes are gauged.

The finding that our participants tend to associate the outcomes of the brain test with one's inborn essence also bears on another large literature, showing that laypeople place undue weight on brain explanations of behavior (Fernandez-Duque, Evans, Christian, & Hodges, 2015; Gruber & Dickerson, 2012; Hook & Farah, 2013; Hopkins et al., 2016; McCabe & Castel, 2008; Michael, Newman, Vuorre, Cumming, & Garry, 2013; Minahan & Siedlecki, 2016; Rhodes, Rodriguez, & Shah, 2014; Schweitzer, Baker, & Risko, 2013; Weisberg et al., 2018; Weisberg et al., 2008; Weisberg et al., 2015.). The possibility that people believe that brain results reflect their inborn essence offers an explanation for the seductive allure of neuroscience.

Our present results are limited, inasmuch as they obtain from Western Educated Industrial participants who are possibly Rich and Democratic (Henrich, Heine, & Norenzayan, 2010); whether these conclusions would apply elsewhere, to small scale societies, remains to be seen. These limitations are particularly pressing given that past research has found that laypeople's notion of the true self is shaped by their own value judgments (Newman et al., 2014a). The possibility thus arises that the preferences we have unveiled here might be limited by the value judgments of our participants.

These limitations notwithstanding, our conclusions call laypeople's intuitive psychology into question. Laypeople believe that they understand the basic workings of the psyche, and they center these narratives around a single unitary notion of the self. The present results suggest that this understanding of the psyche is inaccurate. Rather than holding a single "true self", people seem to tacitly entertain multiple conflicting notions. These notions might arise from two distinct principles of intuitive psychology—Dualism and Essentialism. Dualism might guide the belief in a good moral core that is independent of the body; the grounding of our essence in the body might arise from Essentialism. Whether Dualism and Essentialism are indeed the causes of these inaccurate narratives awaits further research. It appears, however, that just as the ancient Greeks feared, the psychological stories people tell about themselves might be fundamentally skewed (Berent, 2020).

Acknowledgement

We are indebted to Dr. Rachel Theodore, for her generous advice on various statistical matters over the entire course of this project. We also thank Dr. David DeSteno, for his advice on power analysis, and Dr. Kirstin Laurin, for editorial guidance. The generosity of all these individuals attests to the premise of a good true self.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104100.

References

Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths? *Science (New York, N.Y.), 337*(6096), 846. https://doi.org/10.1126/science.1219569.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., … Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Psychological and Cognitive Sciences (Report), 113*(17), 4688. https://doi.org/10.1073/pnas.1522070113.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260–268. https://doi.org/10.1177/0146167208327217.

Ben-Shachar, M. S., Makowski, D., & Lüdecke, D. (2020). *Compute and interpret indices of effect size*. CRAN.

Berent, I. (2020). *The blind storyteller: how we reason about human nature*. Oxford University Press.

Berent, I., Barrett, L. F., & Platt, M. (2020). Essentialist biases in reasoning about emotions. *Frontiers In Psychology: Cognitive Science*, 23. https://doi.org/10.3389/fpsyg.2020.562666.

Berent, I., Platt, M., & Sandoboe, G. M. (in press). Empiricism is natural: It arises from dualism and essentialism. In *Oxford Studies in Experimental Philosophy*.

Blok, S., Newman, G., Behr, J., & Rips, L. J. (2001). Inferences about personal identity. In *Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society*.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

Clark, C. J., Winegard, B. M., & Baumeister, R. F. (2019). Forget the folk: Moral responsibility preservation motives and other conditions for compatibilism. *Frontiers in Psychology, 10*(215). https://doi.org/10.3389/fpsyg.2019.00215.

Dasgupta, N. (2004). Implicit Ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research, 17*(2), 143–169. https://doi.org/10.1023/b:sore.0000027407.70241.15.

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology, 74*, 307–316. https://doi.org/10.1016/j.jesp.2017.10.006.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636. https://doi.org/10.1016/j.tics.2017.05.009.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science, 42*(Suppl. 1), 134–160. https://doi.org/10.1111/cogs.12505.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science, 41*(Suppl. 3), 382–402. https://doi.org/10.1111/cogs.12364.

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience, 27*(5), 926–944. https://doi.org/10.1162/jocn_a_00750.

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

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition, 38*(3), 213–244. https://doi.org/10.1016/0010-0277(91)90007-Q.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions: Biological Sciences, 359*(1451), 1775–1785. https://doi.org/10.1098/rstb.2004.1546.

Gruber, D., & Dickerson, J. A. (2012). Persuasive images in popular science: Testing judgments of scientific reasoning and credibility. *Public Understanding of Science, 21*(8), 938–948. https://doi.org/10.1177/0963662512454072.

Gurley, J. R., & Marcus, D. K. (2008). The effects of neuroimaging and brain injury on insanity defenses. *Behavioral Sciences & the Law, 26*(1), 85–97. https://doi.org/10.1002/bsl.797.

Heath, W. P., Stone, J., Darley, J. M., & Grannemann, B. D. (2003). Yes, I did it, but don't blame me: Perceptions of excuse defenses. *Journal of Psychiatry & Law, 31*(2), 187–226. https://doi.org/10.1177/009318530303100204.

Heiphetz, L. (2019). Moral essentialism and generosity among children and adults. *Journal of Experimental Psychology. General, 148*(12), 2077–2090. https://doi.org/10.1037/xge0000587.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science, 41*(3), 744–767. https://doi.org/10.1111/cogs.12354.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Hirschfeld, L. A. (1995). Do children have a theory of race? *Cognition, 54*(2), 209–252.

Hook, C. J., & Farah, M. J. (2013). Look again: Effects of brain images and mind–brain dualism on lay evaluations of research. *Journal of Cognitive Neuroscience, 25*(9), 1397–1405. https://doi.org/10.1162/jocn_a_00407.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition, 155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011.

Keil, F. C. (1986). The acquisition of natural kind and artifact term. In W. Demopoulos, & A. Marras (Eds.), *Language learning and concept acquisition* (pp. 133–153). New Jersey: Ablex: Norwood.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *The American Psychologist, 74*(5), 569–586. https://doi.org/10.1037/amp0000364.

Martin, N. D., Rigoni, D., & Vohs, K. D. (2017). Free will beliefs predict attitudes toward unethical behavior and criminal punishment. *Proceedings of the National Academy of Sciences, 114*(28), 7325–7733. https://doi.org/10.1073/pnas.1702119114.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*(1), 343–352. https://doi.org/10.1016/j.cognition.2007.07.017.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). New York, NY: Cambridge University Press.

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non)persuasive power of a brain image. *Psychonomic Bulletin & Review, 20*(4), 720–725. https://doi.org/10.3758/s13423-013-0391-6.

Minahan, J., & Siedlecki, K. L. (2016). Individual differences in need for cognition influence the evaluation of circular scientific explanations. *Personality and Individual Differences, 99*, 113–117. https://doi.org/10.1016/j.paid.2016.04.074.

Molouki, S., & Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology, 93*, 1–17. https://doi.org/10.1016/j.cogpsych.2016.11.006.

Monterosso, J., Royzman, E. B., & Schwartz, B. (2005). Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics & Behavior, 15*(2), 139–158. https://doi.org/10.1207/s15327019eb1502_4.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584. https://doi.org/10.1080/09515080500264180.

Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction. *Cognition, 133*(2), 502.

Newman, G. E., Bloom, P., & Knobe, J. (2014a). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216. https://doi.org/10.1177/0146167213508791.

Newman, G. E., Bloom, P., & Knobe, J. (2014b). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216. https://doi.org/10.1177/0146167213508791.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*(1), 96–125. https://doi.org/10.1111/cogs.12134.

Newman, G. E., & Keil, F. C. (2008). Where is the essence? Developmental shifts in children's beliefs about internal features. *Child Development, 79*(5), 1344–1356. https://doi.org/10.1111/j.1467-8624.2008.01192.x.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science, 331* (6023), 1401. https://doi.org/10.1126/science.1192931.

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

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The Cognitive science of folk intuitions. *Noûs, 41*(4), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*(5), 729–738. https://doi.org/10.1162/089892900562552.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General, 146*(2), 165–181. https://doi.org/10.1037/xge0000252.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell Publishing.

Rhodes, R. E., Rodriguez, F., & Shah, P. (2014). Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1432–1440. https://doi.org/10.1037/a0036844.

Roskies, A. L., & Nichols, S. (2008). Bringing moral responsibility down to earth. *The Journal of Philosophy, 105*(7), 371–388. https://doi.org/10.5840/jphil2008105737.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language, 25*(3), 346–358. https://doi.org/10.1111/j.1468-0017.2010.01393.x.

Schweitzer, N. J., Baker, D. A., & Risko, E. F. (2013). Fooled by the brain: Re-examining the influence of neuroimages. *Cognition, 129*(3), 501–511. https://doi.org/10.1016/j.cognition.2013.08.009.

Seto, E., & Hicks, J. A. (2016). Disassociating the agent from the self: Undermining belief in free will diminishes true self-knowledge. *Social Psychological and Personality Science, 7*(7), 726–734. https://doi.org/10.1177/1948550616653810.

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences of the United States of America, 110*(40), 15937–15942. https://doi.org/10.1073/pnas.1314075110.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., … Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science, 25*(8), 1563–1570. https://doi.org/10.1177/0956797614534693.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*(2), 255–259. https://doi.org/10.1177/1745691617698146.

Solomon, G. E., Johnson, S. C., Zaitchik, D., & Carey, S. (1996). Like father, like son: Young children's understanding of how and why offspring resemble their parents. *Child Development, 67*(1), 151–171.

Springer, K., & Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development, 62*(4), 767. https://doi.org/10.2307/1131176.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science, 12*(4), 551–560. https://doi.org/10.1177/1745691616689495.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005.

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

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49–54. https://doi.org/10.1111/j.1467-9280.2008.02045.x.

Waxman, S., Medin, D., & Ross, N. (2007). Folkbiological reasoning from a cross-cultural developmental perspective: Early essentialist notions are shaped by cultural beliefs. *Developmental Psychology, 43*(2), 294–308.

Weisberg, D. S., Hopkins, E. J., & Taylor, J. C. V. (2018). People's explanatory preferences for scientific phenomena. *Cognitive Reserach Principles and Implications, 3* (1), 1–14. https://doi.org/10.1186/s41235-018-0135-2.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040.

Weisberg, D. S., Taylor, J. C. V., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision making, 10*(5), 429–441.