Value Judgments and the True Self

Personality and Social Psychology Bulletin
XX(X) 1–14

DOI: 10.1177/0146167213508791
pspb.sagepub.com
SAGE

George E. Newman[1], Paul Bloom[1], and Joshua Knobe[1]

Abstract

The belief that individuals have a "true self" plays an important role in many areas of psychology as well as everyday life. The present studies demonstrate that people have a general tendency to conclude that the true self is fundamentally good—that is, that deep inside every individual, there is something motivating him or her to behave in ways that are virtuous. Study 1 finds that observers are more likely to see a person's true self reflected in behaviors they deem to be morally good than in behaviors they deem to be bad. Study 2 replicates this effect and demonstrates observers' own moral values influence what they judge to be another person's true self. Finally, Study 3 finds that this normative view of the true self is independent of the particular type of mental state (beliefs vs. feelings) that is seen as responsible for an agent's behavior.

Keywords

true self, attribution, moral reasoning, authenticity, positivity bias, psychological essentialism

Received June 18, 2013; revision accepted September 22, 2013

> Instead of always harping on a man's faults, tell him of his virtues. Try to pull him out of his rut of bad habits. Hold up to him his better self, his REAL self that can dare and do and win out!
>
> —Eleanor Porter, Pollyanna

People sometimes explain behavior by appealing to a set of psychological properties that we may describe as a person's "true" (or "authentic," or "real") self. This notion of a true self plays an important role in many areas of psychology. For example, beliefs about the true self have been shown to influence attributions about behavior (Johnson & Boyd, 1995; Johnson, Robinson, & Mitchell, 2004; Landau et al., 2011; Sripada, 2010), assessments of others' lives (Newman, Lockhart, & Keil, 2010), beliefs about the meaning of life (Schlegel, Hicks, Arndt, & King, 2009; Schlegel, Hicks, King, & Arndt, 2011), decision making (Schlegel, Hicks, Davis, Hirsch, & Smith, 2013), and general measures of well-being (e.g., Kernis & Goldman, 2004; Schimel, Arndt, Pyszczynski, & Greenberg, 2001). The notion of a true self is also prevalent in many areas of society—from Polonius's "To thine own self be true," to advertisements for yoga retreats ("Unlock your soul and become your authentic self"), to the standard advice given to job interviewees and nervous adolescents: "Just be yourself."

Past research on the true self has typically examined how people conceive of their own true selves and the role that it plays in self-concept maintenance. One general conclusion from this work is that people tend to see their own true selves as *virtuous*. For example, thinking about the true self leads to increased self-esteem (Andersen & Williams, 1985), and merely activating the concept of the true self reduces defensiveness (Schimel et al., 2001). Moreover, the tendency to view the true self positively may have important downstream consequences for how people create meaning in their lives (Schlegel et al., 2009, 2011) and make decisions. For example, Schlegel et al. (2013) found that increasing perceived knowledge of the true self boosts satisfaction with major life decisions. In sum, there appears to be a strong normative component to people's conception of their own true selves—people associate the true self with positive characteristics and seem to experience a number of psychological benefits when they feel "in touch" with that concept.

At first blush, it may not seem that surprising that people think of their true selves as fundamentally good. After all, there is a great deal of research suggesting that people tend to be generally positive about many things regarding the self. People overestimate their own knowledge and abilities (Massey, Simmons, & Armor, 2011; Rozenblit & Keil, 2002; West & Stanovich, 1997) and routinely believe their abilities and outcomes to be "above average" (e.g., Gilovich, 1991; Kruger & Dunning, 1999). In addition, people tend to attribute personal accomplishments to the self while downplaying the role of situational factors (e.g., Jones & Nisbett, 1971). Therefore, it could be that the effects observed for how people reason about their own true selves are simply an example of a much broader phenomenon whereby people tend to view themselves in a positive light.

[1]Yale University, New Haven, CT, USA

Corresponding Author:
George Newman, Yale University, 135 Prospect Ave., New Haven, CT 06511, USA.
Email:george.newman@yale.edu

Another possibility, however, is that the positivity associated with the true self reflects something more fundamental about people's understanding of that concept. Independent of anything about people's well-documented tendency to see themselves in a positive light, it might be that people have an intuition that, deep down, inside every individual, there is something motivating him or her to behave in ways that are fundamentally good. Thus, even when people see clearly that certain agents have many bad traits, they might maintain that there is something within these agents—some deeper, hidden element—that is calling them to lead a better life.

Why might this be the case? Although existing studies have not looked directly at people's intuitions about the true self, research on related topics has documented a number of instances in which people's values inform their beliefs about phenomena that appear to be decidedly nonnormative in nature. For example, people's moral judgments appear to affect their intuitions about whether an agent acted "intentionally" (e.g., Cushman & Mele, 2008; Leslie, Knobe, & Cohen, 2006), whether an agent was "in favor" of an outcome (Pettit & Knobe, 2009), whether an agent is "happy" (Phillips, Misenheimer, & Knobe, 2011), and even whether an agent "knew" that an outcome would occur (Beebe & Buckwalter, 2010). In each of these cases, people's ascriptions of psychological states to an agent appear to be influenced by their own value judgments. These patterns are thought to provide evidence for the view that moral considerations actually figure into the most fundamental concepts that people use to make sense of the world (Knobe, 2010). In a similar manner, it may be that people's beliefs about the true self are based in part on their own values and as a result, people show a tendency to conclude that the "true self" reflects whatever they themselves believe to be virtuous.

This tendency would lead people to see their own true selves as good, but it would also lead to a parallel effect in their judgments about others. To date, however, past research on the true self has mainly focused on "first-person" beliefs—that is, how people normatively evaluate their own true selves. Less attention has been devoted to examining people's "third-person" beliefs—that is, how people think about and evaluate the true selves of others. Therefore, the goal of the present studies was to test whether the normative bias associated with the true self extends to people's attributions about others. We examined this across three studies.

Study 1 examined whether people view others' true selves as fundamentally good by testing whether they are more likely to see the true self reflected in changes in behavior they deem to be morally good than in changes they deem to be morally bad.

Study 2 went on to test another prediction that follows from the proposal above. Specifically, if people are predisposed to believe that the true self is fundamentally good, then what constitutes another person's true self should depend on participants' own values—that is, their own conception of right and wrong. Therefore, we predict that, for the same scenario, individual differences in moral values should predict people's true self-attributions.

Study 3 explored these patterns in greater detail. Within philosophy, it has long been claimed that an agent's true self can be identified with the more reflective aspects of the mind (Aristotle, 1985/350 B.C.; Frankfurt, 1971). However, there is also research that points to the opposite view—that people identify the true self with an agent's urges and emotions, rather than the agent's outward behaviors (Johnson et al., 2004). The goal of Study 3 was to examine these potentially different aspects of the self (beliefs vs. feelings) to determine whether the normative view of the true self that we advance here is independent of the particular type of mental state that is seen as responsible for an individual's behavior.

Study 1

In the first study, participants read about a series of individuals who underwent changes in their behaviors and/or beliefs. Some of the agents changed from bad to good, while others changed from good to bad. We then asked participants to use a forced-choice measure to indicate whether they felt that this change reflected the agent's "true self," "surface self," or "none of the above." They also used a rating scale to indicate the extent to which the new behavior/belief was "true to the deepest, most essential aspects of (the agent's) being." Overall, we predicted an asymmetry in judgments, such that participants would be more likely to see the true self reflected in changes they thought were morally good than in changes they thought were morally bad.

To provide a set of "control" vignettes, participants also read about changes along preference dimensions (e.g., preferring dogs to cats). We hypothesized that in these cases, participants would not show an asymmetry in their judgments about the true self because these types of preferences should be seen as too trivial to be diagnostic of the person's underlying essence.

Method

Participants. One hundred and thirty adults (M_{age} = 37.0, 72% female) were recruited through a service that hosts online studies for academic purposes. Participants were compensated via entrance into lotteries for gift certificates.

Stimuli. Stimuli for this study consisted of 12 pairs of vignettes. Each vignette began with the following sentence: "Imagine an individual named ___. ___ is different from you in almost every way—he has a different occupation and prefers different things than you." We did this to reduce the likelihood that participants would infer similarity to the targets, which could complicate interpretation of our findings.

Then, participants read that this individual underwent a change in behavior and/or beliefs. Each vignette had the

structure that the person used to engage in Behavior *X* and now engaged in Behavior *Y.* The direction of the change (from *X* to *Y* vs. *Y* to *X*) was counterbalanced between conditions. Four of the scenarios were changes which we hypothesized participants would view as "good" (i.e., from morally bad behavior to morally good behavior), four were changes which we hypothesized they would view as "bad" (i.e., from morally good behavior to morally bad behavior), and four were neutral (changes in preferences; see Appendix A for all vignettes used in this study).

At the end of each vignette, participants responded to a forced-choice item in which they were asked to specify which aspect of the target's personality they believed to be responsible for the change in behavior. For example, participants read, "In your opinion, what aspect of Omar's personality caused him to treat ethnic minorities with respect (mistreat ethnic minorities)?" The three forced-choice options included (a) "His 'true self' (the deepest, most essential aspect of his being)," (b), "His 'surface self' (the things that he leaned from society or others)," (c) "None of the above." If participants selected the third option, they were provided with a blank space which offered them the opportunity to write a few sentences explaining why they thought the change occurred.

A second rating scale asked about the target's current behavior (following the change) and assessed the degree to which people believed the new behavior/belief reflected the person's true self. For example, participants read, "Now that Omar treats minorities with respect (mistreats minorities), to what extent is he being true to the deepest, most essential aspects of his being?" Participants responded using a 9-point scale with *not at all* and *very much so* as end points.

At the end of the study, participants also reported basic demographic information (age and gender), political orientation (using a binary-choice measure between liberal and conservative) as well as their preferences for the four neutral items. Responses to the preference items were made using 5-point scales with, for example, *strongly prefer dogs* and *strongly prefer cats* as the end points and *no preference* as the midpoint.

Procedure. This study used a mixed-model design such that each participant read four "good" vignettes, four "bad" vignettes, and four "neutral" vignettes. However, the corresponding matched-item pairs were always presented between participants. This produced a 3 (vignette type: good, bad, neutral) × 2 (block 1 vs. block 2) mixed-model design. The order in which the item was presented was randomized for each participant.

Results

Forced-choice items. To assess the overall pattern of results, we recoded the forced-choice item as a binary response with "true self" response as "1" and the "surface self" or "other" responses as "0." We then summed across the morally good, morally bad, and neutral items to produce three scores for each participant, ranging from 0 (no endorsement of the true self) to 4 (endorsement of the true self across all items of that type). A 3 (vignette type: good, bad, neutral) × 2 (block 1 vs. block 2) mixed-model ANOVA revealed a significant main effect of vignette type, $F(2, 127) = 39.92$, $p < .001$, but no effect of block type and no interaction (both Fs < 1). As predicted, participants were more likely to report that the true self had caused the change in behavior/beliefs when the change was morally good ($M = 2.19$, $SE = .12$) compared with when the change was either morally bad ($M = 1.22$, $SE = .12$), $t(128) = 5.98$, $p < .001$, or the change was neutral ($M = .96$, $SE = .10$), $t(128) = 8.94$, $p < .001$.

Comparison of the between-subjects effects revealed a similar pattern. We performed a chi-square analysis (with "true self," "surface self," and "other" as the three potential choices) for each of the 12 items. As seen in Table 1, for most of the moralized behaviors (except for the boyfriend and mistreatment of employees items), participants were significantly more likely to think that the true self was responsible for the morally good changes than for the morally bad changes.

A series of binomial tests examined the choices within each cell (see Table 1). For five of the eight of the morally good behaviors, there were significantly more "true self" responses than "surface self" responses—although for one of the morally good behaviors (respect/mistreat employees), there were significantly more "surface self" responses than "true self" responses. For five of the eight of the morally bad behaviors, there were significantly more "surface self" responses than "true self" responses. And, for six out of eight neutral options, there were significantly more "surface self" responses than "true self" responses. In general, the overall pattern was that participants tended to report that the true self was responsible for morally good changes and that the surface self was responsible for morally bad changes and neutral (nonmoral) changes. (Endorsement of "other" explanations was relatively infrequent across items, and an examination of the write-in responses revealed no systematic patterns.)

True self-rating. We performed a similar set of analysis on the item assessing the degree to which people thought that the target was now behaving in a manner that reflected the true self. A 3 (vignette type: good, bad, neutral) × 2 (block 1 vs. block 2) mixed-model ANOVA revealed a significant main effect of vignette type, $F(2, 127) = 31.01$, $p < .001$, but no effect of block type and no interaction (both Fs < 1). As predicted, participants were significantly more likely to report that the behavior was consistent with the true self when the behavior was morally good ($M = 6.32$, $SE = .13$) compared with when the behavior was either morally bad ($M = 4.86$, $SE = .16$), $t(128) = 6.41$, $p < .001$. Comparison with the neutral vignettes ($M = 5.50$, $SE = .12$) indicated that morally good changes were seen as more revealing of the true self, $t(128) = 5.88$, $p < .001$, while morally bad changes were seen as less revealing of the true self, $t(128) = 3.52$, $p = .001$.

Table 1. Results From Study 1: Frequency of Forced-Choice Responses and True Self-Ratings for Each of the 12 Item Pairs.

	Forced-choice measure (*N* choosing each option)						True self-ratings			
	Good behavior			Bad behavior			Good	Bad		
Item	True self	Surface self	Other	True self	Surface self	Other	χ^2	*t* test		
Moral behaviors										
Honest/corrupt officer	36[a]	17	8	21	35[a]	10	.006	6.45	4.61	.001
Ethical/unethical businessman	39[a]	13	9	21	38[a]	6	.001	6.21	5.00	.002
Treatment of minorities	34	26	6	15	37[a]	10	.006	6.15	4.58	.001
Teetotaler/alcoholic	40[a]	21	5	18	34[a]	10	.002	6.52	4.78	.001
Against/support terrorism	37[a]	17	8	12	47[a]	6	.001	6.68	4.95	.001
Caring/deadbeat father	44[a]	17	6	20	28	14	.001	6.73	4.75	.001
Respect/mistreat employees	23	35[a]	9	20	27	16	.21	5.70	4.83	.02
Good/bad boyfriend	27	25	10	28	30	10	.63	6.15	5.37	.06
	Behavior 1			Behavior 2						
Neutral behaviors										
Mac/ PC computers	7	47[a]	12	10	33[a]	18	.13	5.47	5.40	.84
Country/city	31	26	10	14	33[a]	14	.02	6.06	5.50	.10
Dogs/cats	21	32[a]	14	19	26	17	.66	5.80	5.66	.65
Football/baseball	9	45[a]	11	13	33[a]	16	.18	4.97	5.19	.52

[a]Indicates a significant binomial test result comparing the frequency of "true self" and "surface self" choices.

A series of *t* tests compared the morally good and morally bad conditions for each vignette. As seen in Table 1, for seven of the eight moralized behaviors, participants were significantly more likely to agree that morally good behavior reflected the person's true self compared with the paired morally bad behavior (the exception was the "boyfriend" item, which was only marginally significant, $p = .06$). In contrast, no significant differences were observed across any of the four neutral pairs.

Neutral items and preferences. As noted above, for the preference items, participants showed a general pattern to report that changes in behavior were caused by the "surface self" and did not reflect the "true self" (e.g., ratings were all near the midpoint of "5," as seen in Table 1). To assess whether there were differences in endorsement of the true self based on the strength of people's own preferences, we performed two sets of analyses. The first looked at the raw correlation between the nonmoral preference ratings (taken at the end of the study) and the true self-ratings for each of the neutral items (split by condition). If there is a relationship between these factors, one would predict, for example, that the strength of preference (e.g., strongly preferring dogs over cats) should be positively correlated with the belief that preferring dogs reflects the true self and negatively correlated with the belief that preferring cats reflects the true self. In fact, none of these correlations approached significance (all ps > .3, all rs < .11). A second analysis compared only participants who responded using the end points of the personal preferences measures. We performed a series of one-way ANOVAs treating the preference (e.g., strongly prefer dogs vs. strongly prefer cats) as an independent variable and the true self-rating for that item as a dependent variable. No significant differences were found along any comparisons of this type (all Fs < 1).

Differences in political orientation. The mixed-model ANOVAs reported above were also conducted with political orientation as a factor. There was no interaction or main effect of political orientation on either the forced-choice measure or the rating scale measure (all Fs < 1).

Discussion

Overall, we observed that participants were more likely to attribute morally good changes to the true self, whereas morally bad and neutral (nonmoral) changes were more often attributed to the surface self (i.e., the influence of others or the environment). This pattern was also replicated using a converging rating scale that explicitly asked about the degree to which the target was "true to the deepest, most essential aspects of their being." Taken together, these results suggest that people are predisposed to posit a true self for other people that aligns with their own normative values.

Study 2

The results of Study 1 were consistent with the prediction that people's view of the true self is guided by their own normative values. The goal of Study 2 was to test a more specific prediction that follows from this proposal—namely, if people are predisposed to believe that the true self is fundamentally good, then what constitutes "good" behavior should depend on participants' own view of right and wrong. In other words, individual differences in moral values should predict differences in people's beliefs about the nature of others' true selves.

To test this, we moved to a different design in which all participants evaluated the same scenarios. Instead of manipulating the moral valence of the behavior experimentally, we instead relied on a natural division between liberal and conservative participants. In this study, we again told participants about a series of individuals who underwent a change in behavior/beliefs. However, drawing from research on differences between liberal and conservative moral values (e.g., Graham, Haidt, & Nosek, 2009), we designed the stimuli such that half of the items described a person who changed in a direction that conservatives would be especially likely to regard as good, while the other half described a person who changed in a direction that liberals would be especially likely to regard as good. We then asked about the extent to which the change reflected the emergence of the person's true self. If beliefs about others' true selves are in fact guided by participants' own values, then we should observe an interaction effect such that for the "conservative" items, conservative participants should be more likely than liberals to agree that the person's true self emerged, while for the "liberal" items, liberal participants should be more likely than conservatives to agree that the person's true self emerged.

Note that this design has the additional feature of addressing a potential alternative explanation for the results observed in Study 1. In particular, it may be that there were additional factors that were confounded across the good versus bad behaviors presented in the previous study. For example, many of the morally bad behaviors presented in Study 1 may be destructive to one's health and well-being. Therefore, participants may have conceived of the true self as more of a "survival instinct," rather than as a deeper layer of the person's identity. The design used in Study 2 addressed this potential confound because all participants evaluated the same scenario, and, therefore, "moral valence" was hypothesized to vary across participants, rather than across experimental manipulations.

Method

Participants. A new group of 201 adults (M_{age} = 38.8, 67% female) were recruited through the same online service and were compensated via entrance into lotteries for gift certificates.

Stimuli and procedure. Stimuli for this study consisted of eight vignettes. As in Study 1, each vignette was similar in that it described an individual who underwent a change in behavior/beliefs. Following each vignette, participants were asked to rate the degree to which the change resulted from the emergence of the person's true self (e.g., *At his very essence, there was always something deep within Jim, calling him to stop having sex with men, and then his true self emerged*). Participants responded using a slider bar with "strongly disagree" and "strongly agree" as end points. The corresponding numerical values were 0 and 703 (although no numerical values were visible to participants).

All participants completed the same eight items. However, half of the items were behavioral changes that we predicted that conservative participants would find especially good (homosexuality to heterosexuality, unpatriotic to patriotic, atheist to religious, promiscuous to monogamous), while the other half of the items were behavioral changes that we predicted that liberals would find especially good (deny global warming to supporting the environment, sexist to egalitarian, greedy to generous, and vandalizing abortion clinics to not vandalizing abortion clinics; see Appendix B). Thus, this study used a 2 × 2 mixed-model design with political orientation (liberal vs. conservative) as a between-subjects factor and item type (liberal vs. conservative) as a within-subjects factor. The order in which each item was presented was randomized for each participant.

At the end of the study, all participants were asked to indicate their political orientation using a forced choice between liberal and conservative.

Results

The conservative and liberal items formed reliable scales (α = .81 and .85, respectively). Therefore, for each participant, we averaged the four conservative items and the four liberal items to produce two scales for each participant. We conducted a 2 (political orientation: liberal vs. conservative) × 2 (item type: liberal vs. conservative) mixed-model ANOVA, which revealed a significant interaction between political orientation and item type, $F(1, 199) = 8.44$, $p = .004$. Comparison of the simple effects revealed that conservative participants were more likely to agree that the behavioral change resulted from the emergence of the person's true self for the conservative items ($M = 348.15$, $SD = 150.93$) than for the liberal items ($M = 324.00$, $SD = 138.99$), $t(79) = 2.79$, $p = .04$. Conversely, liberal participants were more likely to agree that the behavioral change resulted from the emergence of the person's true self for the liberal items ($M = 305.16$, $SD = 168.05$) than for the conservative items ($M = 282.89$, $SD = 158.28$), $t(120) = 2.12$, $p = .036$ (see Figure 1).

In addition, we observed a main effect of political orientation such that conservative participants ($M = 336.08$, $SE = 16.37$) gave significantly higher ratings overall than liberal

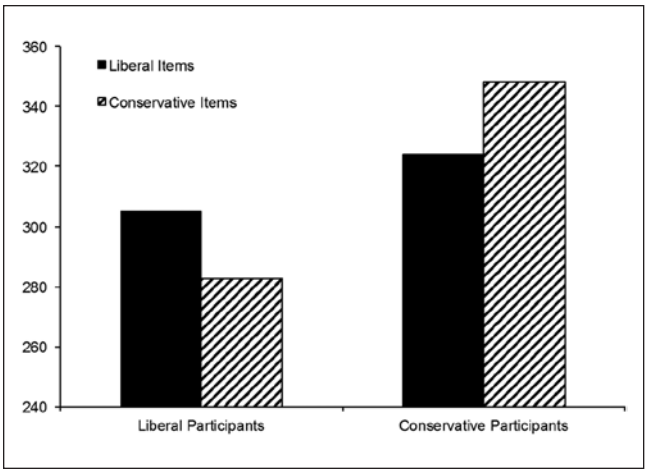

Figure 1. Results from Study 2—Belief that the change in behavior was caused by the emergence of the person's "true self."

participants (M = 294.02, SE = 13.28), $F(1, 199) = 3.99$, p = .047. No other main effects were observed ($F < 1$).

Discussion

The results of this study were again consistent with the notion that people are disposed to endorse a morally virtuous conception of the "true self." Critically, however, this effect was dependent on participants' own moral values and, therefore, systematically varied based on individual differences in political orientation. One result that was not predicted was the main effect for conservatives to provide higher ratings throughout. Based on the results from Study 1, this effect does not appear to be due to a general tendency for conservatives to endorse the true self more than liberals. Rather, this result is interpretable if one assumes that our "conservative" items tested moral values that are more specific to conservatives, whereas our "liberal" items tested values that are held more generally (e.g., liberals and conservatives are reluctant to endorse sexism and being selfish; see Graham et al., 2009).

A second issue is that the absolute level of agreement with the main dependent measure tended to be lower in this study than in Study 1. Perhaps the best explanation for this difference is that the two studies actually asked slightly different questions. Specifically, in Study 1 participants were asked to indicate the extent to which they thought the new behavior (following the behavior change) reflected the person's true self, while in Study 2, participants were instead asked the extent to which "there was always something deep within [Jim], calling him to . . . " In other words, the dependent measure used in Study 2 asked about the existence of an "opposing" true self (i.e., one that was different from all outward behavior) that was present all along. In many respects, this is more extreme test of the true self-belief and, therefore, it is perhaps not that surprising that on average, agreement with this item tended to be lower. We should also stress that the primary effect of interest in this study was the difference between liberal and conservative participants, which provides evidence for a normative conception of the true self.

Study 3

Study 3 examined how beliefs about the true self map onto distinctions between beliefs versus feelings. One view of the true self that has been advanced in the philosophical literature is that the true self is the most clearly reflected in rational deliberation (Aristotle 1985/350 B.C.E.). For example, consider a person who is fighting an addiction to heroin. She might have a continual craving for more heroin, but if she gives in to this craving, this view would say that by doing heroin she is not being true to herself and rather is betraying what she values the most (Frankfurt, 1971). Alternatively, one could imagine a different conception, which maintains that the true self is reflected when people are overcome with a particular emotion or desire. For example, within the work of novelists and poets, there has been a long tradition that points to the opposite view, identifying an agent's true self with precisely those urges and emotions that are only revealed when the agent casts away all deliberation and reflection (e.g., de Sade, 1791/2008; also see Johnson et al., 2004).

Given that the distinction between more (less) deliberative thinking seems to be central to previous theorizing about the true self, we wanted to directly contrast the normative effects observed in the previous studies with any potential differences between beliefs versus feelings. To accomplish this, we asked participants to imagine an individual who had a belief that pulled in one direction (e.g., believing that homosexuality is immoral) but a feeling that pulled in an opposite direction (e.g., an attraction toward people of the same sex). We then asked which of these mental states (the belief or the feeling) reflected the agent's "true self." Following the logic of Study 2, we predicted that individual differences in values (e.g., beliefs about the moral acceptability of homosexuality) would predict whether participants saw either the belief or the feeling as part of the agent's true self. Specifically, in this particular case, liberals more so than conservatives should think that the agent's attraction toward the same sex reflects the person's true self because liberals (vs. conservatives) are more likely to view homosexuality as morally acceptable (e.g., Inbar, Pizarro, & Bloom, 2012).

We also presented an analogous case to a second group of participants in which the beliefs and feelings were reversed. In this scenario, the agent had a belief that homosexuality was perfectly acceptable but had a negative feeling toward the thought of same-sex couples. Again, a normative conception of the true self predicts a difference based on participants' own moral values. Specifically, in this case, conservatives should be more likely than liberals to think that the agent's negative feeling toward same-sex couples reflects the true self. Thus, by crossing the type of mental

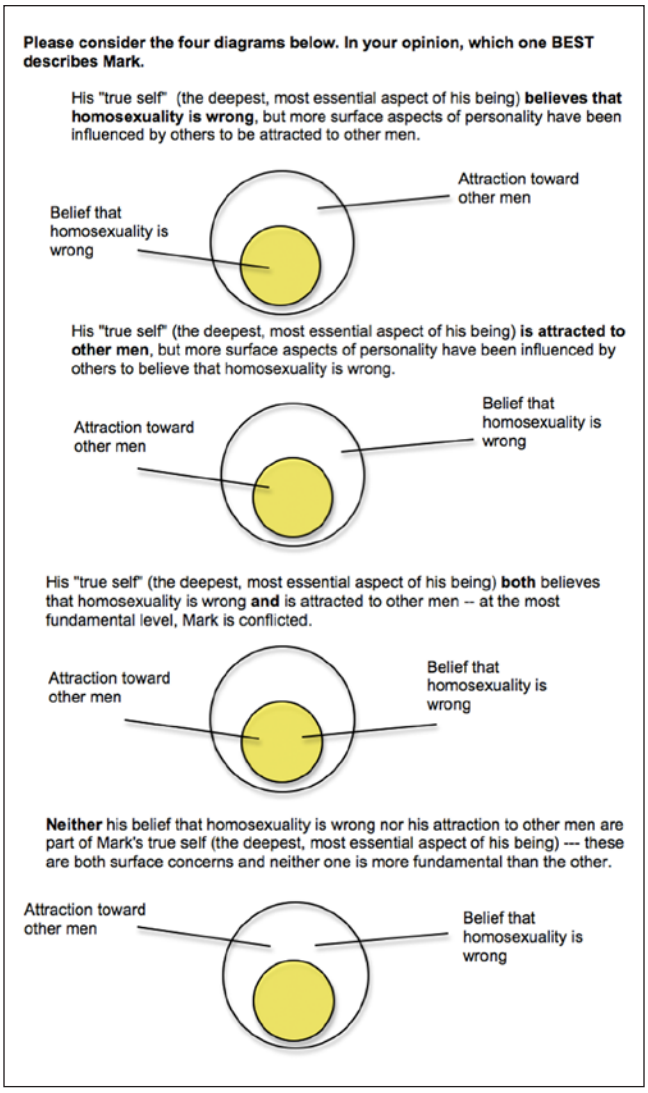

Figure 2. Stimuli presented to participants in one condition of Study 3 (belief = homosexuality immoral, feeling = attraction toward same sex).

state (belief vs. feeling) with the moral valence associated with this state, this methodology allowed us to assess whether the normative effects observed in the previous studies exist independently of the particular type of mental state in question (belief vs. feeling).

In addition, we measured beliefs about the true self in a new way. In this study, participants inspected a series of diagrams that depicted "models" of the agent's personality (see Figure 2). The diagrams were concentric circles that showed either the belief as part of the true self and the feeling as more peripheral, the feeling as part of the true self and the belief as more peripheral, both the feeling and belief as part of the true self, or neither of these as part of the true self. This measure was also accompanied by rating scales, which (as in previous studies) asked about the degree to which the belief and the feelings were part of the person's true self. Together, these measures provided a converging means of assessing how values shape beliefs about the true self and the extent to which this normative effect is independent of additional beliefs about the role of beliefs versus feelings.

Method

Participants. Two-hundred and one adults (M_{age} = 37.2, 64% female) were recruited through the same online service and were compensated via entrance into lotteries for gift certificates.

Stimuli and procedure. Stimuli for this study consisted of two different vignettes that were presented between-subjects. Both vignettes described an individual who had a belief that pulled in one direction and an opposing feeling/desire that pulled in the opposite direction. Specifically, one vignette described an individual who believed that homosexuality was wrong but had an attraction toward other men. This vignette read as follows:

> Mark is an evangelical Christian. He believes that homosexuality is morally wrong. In fact, Mark now leads a seminar in which he coaches homosexuals about techniques they can use to resist their attraction to people of the same-sex.
>
> However, Mark himself is attracted to other men. He openly acknowledges this to other people and discusses it as part of his own personal struggle.

Conversely, the other vignette described an individual who believed that homosexuality was perfectly acceptable but experienced a negative feeling when thinking of same-sex couples. This vignette read as follows:

> Mark is a secular humanist. He believes that homosexuality is perfectly acceptable. In fact, Mark leads a seminar in which he coaches people about techniques they can use to resist their negative feelings about people who are attracted to the same sex.
>
> However, Mark himself has a negative feeling about thought of same-sex couples. He openly acknowledges this to other people and discusses it as part of his own personal struggle.

Participants then completed two manipulation checks that asked them to indicate what Mark believed (forced-choice between "homosexuality is wrong" and "homosexuality is permissible") and what Mark experienced (forced-choice between "attraction toward the same sex" and "negative feelings about people who are attracted to the same sex"). Thirty-eight adults did not pass this manipulation check, which was defined as correctly identifying what the actor believed and what they felt. Given that this was the critical factor that differed across conditions, participants who failed this check were not included in subsequent analyses.

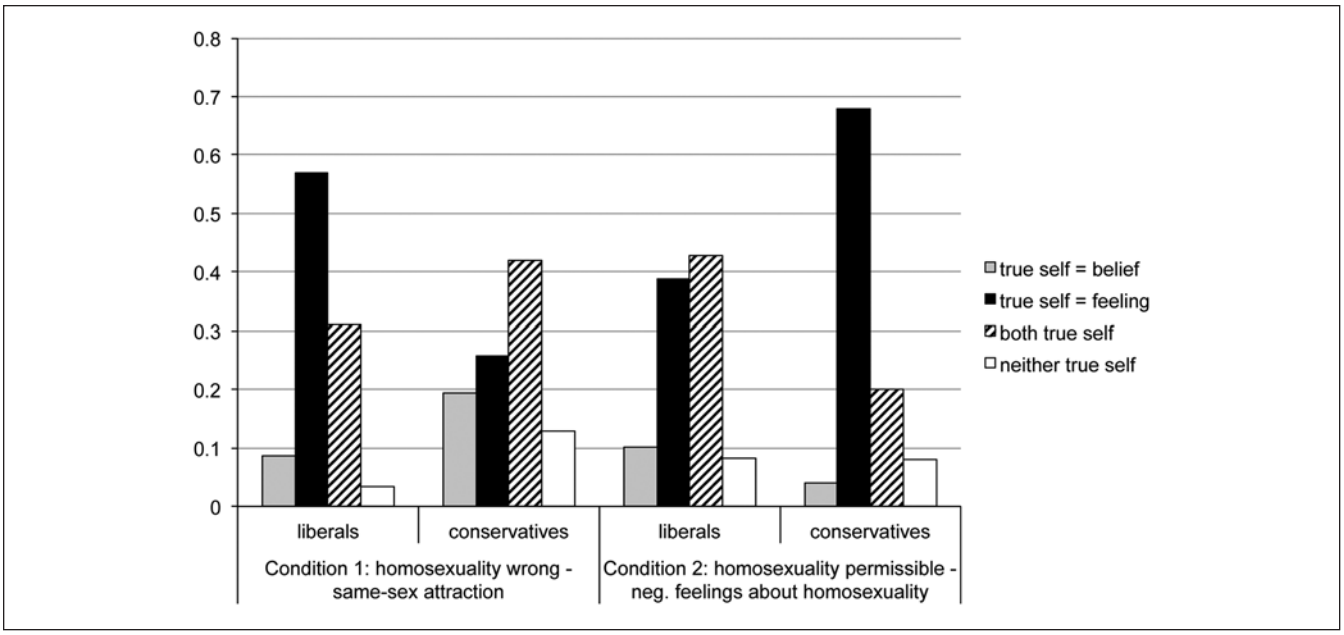

Figure 3. Proportion of liberal and conservative participants who selected each of the models in both conditions of Study 3.

Following, participants were shown a series of four diagrams (see Figure 3) and were asked to choose the one that they thought best represented Mark's personality. The four "models" were as follows: (a) the belief (homosexuality is wrong/permissible) was Mark's true self and the feeling (attraction toward same sex/negative feelings toward homosexuality) was more peripheral, (b) the feeling was the true self and the belief was more peripheral, (c) both the belief and feeling were part of the true self, and (d) neither the belief nor the feeling were part of the true self. Finally, participants responded to two items that asked them to imagine that Mark no longer had the same belief (but had the same feeling) and a second item that asked them to imagine that Mark no longer had the same feeling (but had the same belief). In both cases, they indicated the extent to which Mark would "still be true to the deepest, most essential aspects of his being" (1 = not at all, 9 = very much so). At the end of the study, participants reported basic demographic information including political orientation using the same binary measure as in previous studies.

Results

Personality models. Responses on the forced-choice measure were consistent with the results of the previous studies. As seen in Figure 3, there was a strong overall tendency for participants to regard Mark's feelings as part of his true self. However, participants' intuitions about the true self also showed the predicted impact of individual differences, with participants being more likely to regard Mark's psychological states as part of his true self when those states fit with their own values.

Specifically, for the condition where Mark believed that homosexuality was wrong, but was attracted toward other men, the majority of liberal respondents (57%) said that the Mark's attraction toward other men was his true self while his belief was peripheral. In contrast, only 26% of the conservative participants selected this option and the majority (42%) thought that the belief that homosexuality was wrong and his attraction toward other men were part of the true self. This difference was confirmed via a chi-square analysis comparing conservative and liberal responding across all four items, $\chi^2 = 9.49$, $p = .02$. For the scenario in which Mark believed that homosexuality was permissible, but had a negative feeling toward homosexuality, the pattern was reversed. Here, the majority of conservatives (68%) responded that Mark's true self was his negative feeling toward homosexuality while his belief was more peripheral. In contrast, only 38% of the liberal participants selected that option, and the majority of liberals (43%) responded that the feeling and belief were part of Mark's true self. This difference was marginally significant, $\chi^2 = 6.15$, $p = .10$.

True self-ratings. Results for these measures are depicted in Figure 4. We conducted a 2 (political orientation: liberal vs. conservative) × 2 (scenario type) × 2 (item type: only belief remains vs. only feeling remains) mixed-model ANOVA. Consistent with the patterns above, this analysis revealed a significant three-way interaction, $F(1, 157) = 11.92$, $p < .001$.

We then examined each scenario individually. For the scenario in which Mark believed that homosexuality was wrong, but was attracted toward men, we observed an interaction between political orientation and item type, $F(1, 85) = 8.99$, $p = .004$. Liberals were more likely than conservatives to

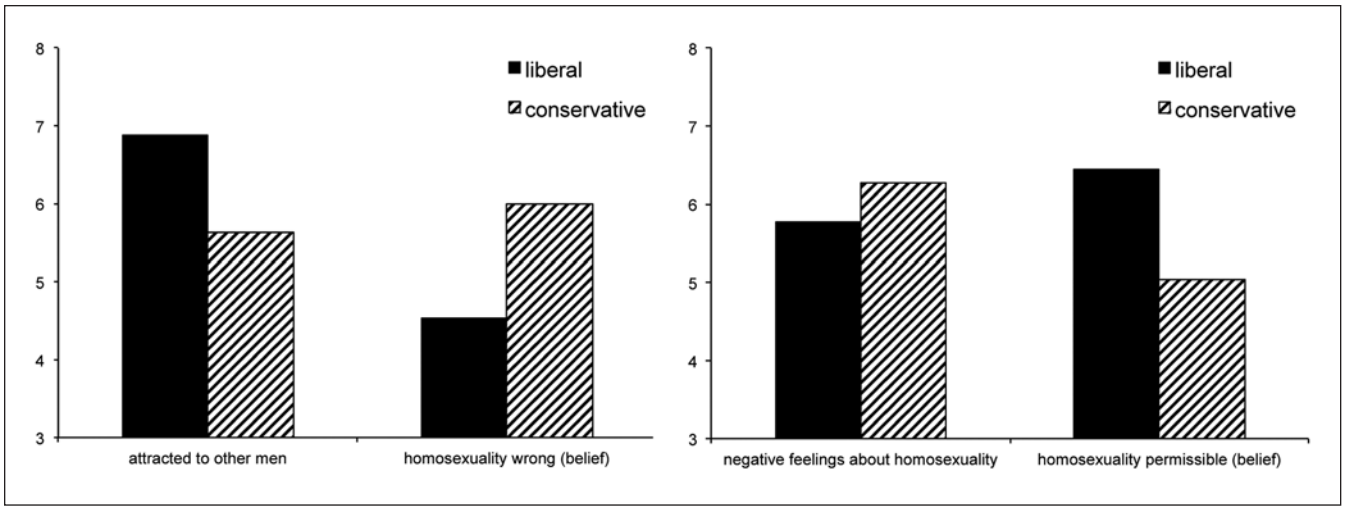

Figure 4. Study 3 results—Agreement that agent would be "true to the deepest most essential aspects of his being" if he possessed various traits (as a function of political orientation).

believe that Mark would still be true to his essence if he was attracted to other men (*M*s = 6.88 and 5.63, respectively). Conversely, conservatives were more likely than liberals to believe that Mark would be true to his essence if he believed that homosexuality was wrong (*M*s = 5.97 and 4.54, respectively). For the scenario in which Mark believed that homosexuality was permissible, but has a negative feeling toward the idea of homosexuality, we found the opposite pattern, $F(1, 72) = 3.76$, $p = .056$. Conservatives were more likely than liberals to believe that Mark would be true to his essence if he had negative feelings toward homosexuality (*M*s = 6.28 and 5.78, respectively), while liberals were more likely than conservatives to believe that Mark would be true to his essence if he believed that homosexuality was permissible (*M*s = 6.45 and 5.04, respectively).

Finally, this 2 × 2 × 2 ANOVA revealed a marginal main effect of item type, where overall participants reported that feelings ($M = 6.14$, $SE = .21$) were more consistent with the "true self" compared with beliefs ($M = 5.50$, $SE = .21$), $F(1, 157) = 3.74$, $p = .055$. This result is consistent with the data from the forced-choice item where people appear to have a general belief that feelings are more representative of the true self. Importantly, however, that tendency appears to be distinct from the normative view to see the true self as virtuous.

Discussion

The results from this study were interesting for multiple reasons. First, we replicated the normative effect found in previous studies where participants are more inclined to see a psychological state as part of the true self when they think that it is good than when they think it is bad.

Second, we found that there seems to be a general tendency for people to see feelings as part of the true self than to see beliefs as part of the true self. This result is consistent with past research (Johnson et al., 2004), which has also found that feelings tend to be seen as more diagnostic of the true self than outward behaviors. Together, these findings indicate that people associate the true self with precisely those mental states that are "unwanted" (compared with those that are more deliberate), suggesting that people conceive of the true self as an entity that is separate from, and perhaps immune to, one's explicit goals and beliefs.

Finally, it is worth noting that this study used a similar rating scale as in Study 1, and the responses were also roughly at the midpoint of the scale, suggesting that the absolute differences observed in Study 2 reflected a difference in the type of question that was asked.

This pattern of results now makes it possible to explain why philosophers might have been drawn to the view that the true self was constituted by moral beliefs rather than feelings. In the case where a heroin addict feels an urge to get another fix but believes that she should refrain, it might indeed seem intuitive that her belief is part of the true self but her urge is not (Frankfurt, 1971). However, this intuition arises only because people think that the belief is good and the urge is bad. If we now reverse the story—with the addict believing that he should try heroin but having an urge not to—we would predict that people would have precisely the opposite intuition.

General Discussion

Across three studies, we found that people show a general tendency to conclude that deep inside every individual, there is a "true self" motivating him or her to behave in ways that are virtuous. This finding emerged when participants were asked to explain why agents underwent a change in behavior/beliefs (Study 1), when different participants had different

moral intuitions about the same changes in behavior/beliefs (Study 2), and when the target's beliefs and feelings were contrasted and participants were asked to select a visual diagram that best reflected their intuitions about the true self (Study 3).

Together, these results suggest that previous findings of a positivity bias associated with one's own true self seems to stem from the very nature of the true self-concept. In other words, people not only believe their own true self is good—they expect others' true selves to be good as well. Importantly, however, we find that what counts as "good" depends on perceiver's own personal beliefs about right and wrong. As a result, individual differences in values seem to influence people's beliefs about the nature of the true self (as revealed by Studies 2 and 3).

Scope of the Effect

It is important to note that all participants in these studies were from Western cultures. Existing data suggest that people from Western cultures place a special value on the individual (Choi, Nisbett, & Norenzayan, 1999; Henrich, Heine, & Norenzayan, 2010), and one might therefore worry that the results obtained here merely reflect the idiosyncrasies of one particular cultural context. There is anecdotal evidence to suggest that this intuition holds more broadly. For example, in classical Chinese philosophy, one finds the idea that although certain people may seem callous or selfish on the surface, there is something deep within them drawing them toward a more morally good life (e.g., Mencius, 2009). Nevertheless, further research is needed to determine the extent to which this view of the true self is shared cross-culturally.

We should also stress that these results do not indicate that values alone dictate beliefs about the true self. People may also hold preexisting theories about the types of factors that are likely to shape the true self (e.g., genes, childhood upbringing), the types of forces that interact within the true self (e.g., metadesires and willpower), and the types of situations in which the true self is likely to be revealed (e.g., when someone is intoxicated). More specifically, the results of Study 3 suggest that, in addition to be influenced by values, people are more inclined to see a mental state as part of the true self when it involves feelings/desires than when it involves beliefs. As noted earlier, this result is consistent with past research (Johnson et al., 2004), which has also found that feelings (vs. behaviors) are seen as more diagnostic of the true self. This result is quite interesting as it suggests that people's lay conception of the true self seems to prioritize mental states that are automatic and perhaps even unconscious over those that are conscious and more deliberate. All else being equal, one might expect exactly the opposite, where things that are under one's control are seen as more defining of the self than things that are not. This general tendency to see "unwanted" mental states as more reflective of the true self seems worthy of further investigation as it suggests that the true self is seen as something that is discovered or emerges rather than as something that can be constructed or willfully defined.

A final issue concerns the role of similarity. Given people's general tendency to overestimate the extent to which others are similar to themselves (Ross, Greene, & House, 1977) and even possess the same knowledge and mental states (Birch & Bloom, 2007), one might wonder whether the effects documented here simply reflect a more general pattern for people to assume that "deep down," others are the same as they are.

It is perhaps useful to distinguish here between two notions of similarity. In one sense, people could assume that others are similar to themselves in most/all respects (similarity to *me*). However, in another sense, people could assume that others are likely to hold same values deep down (similar to *my values*). To illustrate this difference, consider the results of Study 3. It may be the case that participants who are themselves gay are more likely to think that deep down, the actor's "true self" is also gay (similarity to *me*). However, given that it is unlikely that the majority of our "liberal" participants in Study 3 were gay, there also seems to be an effect based on one's personal values about homosexuality. In other words, even among participants who are not themselves gay, those who endorse the value that homosexuality is morally acceptable (vs. morally bad) are more likely to show a tendency to say that this feeling was a part of his true self (similarity to *my values*).

A similar point applies to the vignettes from Studies 1 and 2 (e.g., being an honest police officer, vandalizing abortion clinics, etc.). It may well be the case that people who themselves have a desire to perform these actions would tend to attribute a corresponding desire to the agent's true self (similarity to *me*), but it seems unlikely that such an effect is at the root of the results obtained here. (It is unlikely that a large percentage of our participants themselves had a desire to be honest police officers or that the "conservative" participants themselves had a desire to vandalize abortion clinics). More plausibly, the effects obtained here reflect people's judgments about the degree to which these behaviors are good or bad (similarity to *my values*).

This distinction between different notions of similarity may also have interesting overlaps with classic distinctions between more descriptive aspects of one's identity ("me" self-concept) versus more experiential aspects ("I" self-concept; James, 1890). Past research on "I-sharing" (e.g., Pinel, Long, Laundau, Stanley, & Pyszczynski, 2006) suggests that perceived commonality in one's experience (which may include one's values and normative beliefs) is critical to feelings of interpersonal connectedness, and it may be that this framework also helps to understand people's beliefs about the true self.

Ultimately, then, while there may be an effect of "similarity to me" (an effect that could be investigated in further

research), the present studies suggest that there is also an independent effect of perceived similarity in value judgments that seems to be responsible for the normative effects observed here.

Relation to Attribution Theory

Another issue that arises is an apparent tension between these findings and results from attribution theory. In a typical attribution study, participants are presented with behaviors performed either by themselves or by others and are asked about the extent to which these behaviors are due to factors within the *person* versus the *situation* (e.g., Jones & Nisbett, 1971). Such research typically finds a difference between people's judgments about themselves and about others. For behaviors performed by themselves, participants tend to attribute the good to the person and the bad to the situation, but for behaviors performed by others, they tend to attribute the good to the situation and the bad to the person (for a meta-analysis, see Malle, 2006). The present studies, however, find a seemingly different pattern. Even for judgments about others, people tend to say that the good behaviors reflect the true self while the bad behaviors do not. At least initially, this finding might appear to conflict with the results from classic attribution studies.

It should be noted, however, that the questions posed to participants in these two types of studies are radically different. In the classic attribution studies, participants are presented with a single behavior and are not given any explicit information about why the agent performed it. Their task is to determine, based on this limited information, whether the behavior was due to something about the agent herself or to something about the external situation (for review, see Gilbert & Malone, 1995). By contrast, in the studies conducted here, participants are explicitly told about a psychological state within the agent. (For example, participants in Study 3 were told, "Mark himself is attracted to other men"). The question then is about the status of this psychological state, with participants being asked to determine whether the state belongs to the "true self" or to the "surface self."

Examining the whole pattern of data across these two types of studies, it seems clear that a proper explanation of the results within each type will have to refer to specific aspects of the actual questions posed to participants. Thus, the explanation of the results in the classic attribution studies cannot be that people show an across-the-board tendency to assume the worst about others; it has to involve something specific to the process people use when making inferences about situational causes of individual behaviors. (For example, it could be that bad behaviors tend to show low consensus, and low consensus behaviors are rarely attributed to situational causes; Kelley, 1967). Conversely, the explanation of the results in the present studies cannot be that people show an across-the-board tendency to think the best about others. It has to involve something specific about the process people use when making true self-attributions and determining *which* aspect of the self was responsible for a given behavior.

Why is the True Self Good?

A question now arises about how exactly this explanation should proceed. What in particular is it about the way people understand the true self that makes them see it as good?

One hypothesis is that this effect is best explained in terms of people's *psychological essentialism* (e.g., Bloom, 2010; Gelman, 2003; Medin & Ortony, 1989; Newman & Keil, 2008). A number of studies point to a surprising effect whereby people's value judgments play a role in their attributions of essences. Research in developmental psychology shows that young children show a bias to believe that positive traits, such as being kind or clean, are retained through development while negative traits spontaneously become more positive over time. Interestingly, this seems to extend even to biological traits such as having poor eyesight or a missing finger (Lockhart, Chang, & Story, 2002), suggesting that there is some nascent idea of a deeper cause that is "restoring" traits to a more positive state. Similarly, research in the psychology of concepts suggests that when people represent the essence of a category, they tend to do so in an idealized way that highlights what the category member *should be* like (Barsalou, 1985; Lynch, Coley, & Medin, 2000). For example, people's values can shape their judgments about what it means to be a "true work of art," "true love," or a "true scientist" (Knobe, Prasada, & Newman, 2013). These various studies all seem to be pointing to the same basic conclusion: Even in cases where people acknowledge that an object or category has certain bad properties, they show a certain inclination to think that, deep down at its very essence, the object or category is actually good. Therefore, one hypothesis is that the effect we find for true self-attributions is a manifestation of this more general fact about the nature of people's essentialist reasoning.

However, this, of course, only pushes the problem back one step. If one assumes that the effect of value judgments on true self-attributions is to be explained in terms of an effect of value judgment on essentialist reasoning, one is immediately faced with a new question: Why exactly is there an effect of value judgment on essentialist reasoning? We see two possible types of explanations, corresponding to the traditional distinction between "motivational" and "cognitive" approaches (e.g., Kunda, 1990).

The first possibility is that people are in some way *motivated* to represent essences in this way. People may have an implicit (or perhaps even explicit) motivation to see their own values as stable, essential aspects of the world. Indeed, past research has documented several instances in which people are motivated to endorse abstract values that do not directly benefit them personally (e.g., Jost, Banaji, & Nosek, 2004), and essentialist reasoning in particular may be one

way in which people go about satisfying those motivations. For example, when the perceived stability of social structures is threatened, female (as well as male) participants are more likely to endorse the idea that gender differences in math and science are due to innate biological factors (Brescoll, Uhlmann & Newman, in press). Therefore, it may be that when reflecting on the most essential aspects of others (e.g., the true self), people are predisposed to conceive of those essences as morally good and conforming to their own sense of right and wrong. Put differently, it may be very difficult for people to view things that they regard as immoral as fundamental, unchanging aspects of the world.

A second possibility would be that the impact of value judgments on essentialist reasoning is purely *cognitive*—that is, it might be that the impact of value judgments is not the result of a motivational bias but is instead a more basic aspect of the way people's essentialist reasoning works. Existing work on the role of value judgments in understanding category essences, for example, has not attributed the observed effects to motivation. For example, people's values can shape their judgments about what it means to be a "true scientist" (Knobe et al., 2013), but this does not seem to result from a motivational bias distorting people's judgments. Instead, it seems that these judgments are value-laden through and through, as people show similar patterns of reasoning for novel, unfamiliar categories (Knobe et al., 2013, Experiment 5).

One might then apply a similar approach to understanding people's lay conception of a "true self." On such a view, it is not that people's core capacity for essentialist reasoning is value-free, but this capacity is then distorted by a motivational bias. Rather, the claim is that value judgments actually play a role in the core capacity itself. (Although this suggestion may at first seem a bit counterintuitive, there is growing evidence for a similar approach in other aspects of folk psychology; for a review, see Knobe, 2010.)

To decide between these alternatives, we will need to go beyond work that looks specifically at judgments about the true self. It seems that there is a more general phenomenon involving the nature of people's essentialist reasoning, and future research should explore that phenomenon as a problem in its own right.

Conclusion

The present studies identified an asymmetry in how people conceive of others' true selves—namely, people show a general tendency to conclude that deep inside every individual, there is a true self calling him or her to behave in ways that are virtuous. This work contributes to existing research on the true self by providing evidence that the positivity bias associated with the true self-concept appears to extend to beliefs about others (as well as the self). More broadly, this research offers evidence for an additional wrinkle in people's lay theories about others as it suggests that while people may be quite willing to attribute bad outcomes to factors within other agents, they also show a tendency to conclude that, deep down, those agents have a "true self" that is fundamentally good.

Appendix A

Stimuli Presented in Experiment 1

Morally Good/Bad Vignettes

Al used to be a "deadbeat" dad. In the past, he never showed any real affection for his children and never expressed any interest in his children's lives. Now, however, Al is a very caring and involved father.

Al used to be a very caring and involved father. In the past, he always showed real affection for his children and always expressed interest in his children's lives. Now, however, Al is not a very caring father and is not involved in his children's lives.

Amir lives in a culture that supports terrorism. In the past, Amir supported the idea of terrorism to achieve political goals. Now, however, Amir believes that terrorism is wrong.

Amir lives in a culture that does not support terrorism. In the past, Amir did not support the idea of terrorism. However, now, Amir believes that terrorism is an acceptable way to achieve political goals.

Bill used to mistreat his employees. In the past, he often yelled at them and publicly embarrassed them for minor infractions. Now, however, he never yells at his employees or does anything to publicly embarrass them.

Bill used to treat his employees well. In the past, he never yelled at them or did anything to publicly embarrass them. Now, however, he often yells at them and publicly embarrasses them for minor infractions.

Frank works in an environment that supports dishonest business practices. In the past, he too has participated in dishonest business practices. Now, however, Frank believes that it is wrong to engage in dishonest business practices and only behaves ethically.

Frank works in an environment that supports only honest business practices. In the past, he has not participated in dishonest business practices. Now, however, Frank believes that it is permissible to engage in dishonest business practices and behaves unethically.

Jim used to be an alcoholic. In the past, he never tried to quit drinking and never expressed any interest in trying to quit. Now, however, Jim does not drink any alcohol.

Jim used to be a teetotaler. In the past, he never tried a drink of alcohol and never expressed any interest in drinking. Now, however, Jim is an alcoholic.

Luke used to be a "jerk boyfriend." In the past, he never treated his girlfriend well. Now, however, Luke is an excellent boyfriend and treats his girlfriend with respect and affection.

Luke used to be an excellent boyfriend. In the past, he never treated his girlfriend poorly. Now, however, Luke is a "jerk boyfriend" and never treats his girlfriend with the proper amount of respect and affection.

Omar lives in a culture that oppresses ethnic minorities. In the past, he also mistreated ethnic minorities and never expressed any interest in giving minorities equal rights. Now, however, he treats ethnic minorities with respect and believes that minorities should have equal rights.

Omar lives in a culture that treats all ethnic groups equally. In the past, he also treated ethnic minorities with respect and believed that minorities should have equal rights. Now, however, he mistreats ethnic minorities and does not think that minorities should have equal rights.

Tom is a police officer and works in a station that supports corruption. In the past, he has also participated in police corruption. Now, however, he does not engage in corrupt activities and always conducts himself in an ethical manner.

Tom is a police officer and works in a station that has never supported police corruption. In the past, he never participated in corruption and always behaved ethically. Now, however, Tom engages in corrupt activities and does not conduct himself in an ethical manner.

Neutral Vignettes

Alex used to only have dogs as pets. Now, however, Alex strongly prefers cats and only has them as pets.

Alex used to only have cats as pets. However, now, Alex strongly prefers dogs and only has them as pets.

Ralph used to use PC computers. Now, however, Ralph uses only Mac computers.

Ralph used to use Mac computers. Now, however, Ralph uses only PC computers.

Rob used to believe that baseball was the best sport and only followed baseball. Now, however, Rob only watches football and thinks that football is the best sport.

Rob used to believe that football was the best sport and only followed football. Now, however, Rob only watches baseball and thinks that baseball is the best sport.

Sam used to live in a big city. In the past, he frequently talked about how much he loved living in an urban environment. Now, however, Sam lives in the country and frequently discusses how much he loves living in a rural environment.

Sam used to live in the country. In the past, he frequently talked about how much he loved living in a rural environment. Now, however, Sam lives in a big city and frequently discusses how much he loves living in an urban environment.

Appendix B

Stimuli Presented in Experiment 2

"Good" for Conservatives

Bill used to be unpatriotic and deeply critical of his country. However, now Bill is very patriotic and openly expresses love for his country.

Dave used to be sexually promiscuous and had sex with multiple partners. However, now Dave is monogamous and has stopped being promiscuous.

Frank used to be an atheist and did not believe in God. However, now Frank is very religious and openly expresses his belief in God.

Jim used to be homosexual. However, now Jim is married to a woman and no longer has sex with men.

"Good" for Liberals

Al used to not believe in global warming. However, now, Al believes global warming is a serious problem and thinks that we should immediately stop burning fossil fuels.

Henry used to not pay attention to gender issues in the workplace. However, now, Henry works hard to think about gender issues and supports women in the workplace.

Ned used to vandalize abortion clinics. However, now, Ned thinks that other people's choice should be respected and no longer vandalizes abortion clinics.

Ralph used to make a lot of money and prioritized his financial success above all else. However, now Ralph works in a job where he helps others but does not make a lot of money.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Andersen, S. M., & Williams, M. (1985). Cognitive/affective reactions in the improvement of self-esteem: When thoughts and feelings make a difference. *Journal of Personality and Social Psychology*, *49*, 1086-1097.

Aristotle. (1985). *Nicomachean ethics* (T. Irwin, Trans.). Indianapolis, IN: Hackett. (First published 350 B.C.E.)

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 629-654.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, *25*, 474-498.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382-386.

Bloom, P. (2010). *How pleasure works*. New Work, NY: Basic Books.

Brescoll, V., Uhlmann, E. L., & Newman, G. E. (in press). The effects of system-justifying motivations on endorsement of essentialist explanations for gender differences. *Journal of Personality and Social Psychology*.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*, 47-63.

Cushman, F. A., & Mele, A. (2008). Intentional action: Two and half folk concepts (pp. 171–188). In J. Knobe & S. Nichols (Eds.), *Experimental philosophy*. New York, NY: Oxford University Press.

de Sade, M. (2008). *Justine, or the misfortunes of virtue*. Radford, VA: Wilder. (Original work published 1791)

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5-20.

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

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*, 21.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York, NY: The Free Press.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029-1046.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-135.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012). Disgusting smells cause decreased liking of gay men. *Emotion*, *12*, 23-27.

James, W. (1890). *Principles of psychology*. Toronto, Ontario, Canada: General Publishing Company.

Johnson, J. T., & Boyd, K. R. (1995). Dispositional traits versus the content of experience: Actor/observer differences in judgments of the "authentic self." *Personality and Social Psychology Bulletin*, *21*, 375-383.

Johnson, J. T., Robinson, M. D., & Mitchell, E. B. (2004). Inferences about the authentic self: When do actions say more than mental states. *Journal of Personality and Social Psychology*, *87*, 615-630.

Jones, E., & Nisbett, R. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79-94). Morristown, NJ: General Learning Press.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*, 881-919.

Kelley, H. H. (1967). Attribution theory in social psychology (pp. 192–238). In D. Levine (Ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press.

Kernis, M. H., & Goldman, B. M. (2004). Authenticity, social motivation, and wellbeing. In J. P. Forgas, K. D. Williams & S. Laham (Eds.), *Social motivation: Conscious and unconscious processes* (pp. 210-227). New York, NY: Cambridge University Press.

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

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

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121-1134.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480-498.

Landau, M. J., Vess, M., Arndt, J., Rothschild, Z. K., Sullivan, D., & Atchley, R. A. (2011). Embodied metaphor and the "true" self: Priming entity expansion and protection influences intrinsic self-expressions in self-perceptions and interpersonal behavior. *Journal of Experimental Social Psychology*, *47*, 79-87.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: "Theory of mind" and moral judgment. *Psychological Science*, *17*, 421-427.

Lockhart, K., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism? *Child Development*, *73*, 1408-1430.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded-category structure among tree experts and novices. *Memory & Cognition*, *28*, 41-50.

Malle, B. F. (2006). The actor-observer asymmetry in causal attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*, 895-919.

Massey, M., Simmons, J., & Armor, D. A. (2011). Hope over experience: Desirability and the persistence of optimism. *Psychological Science*, *22*, 274-281.

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

Mencius. (2009). *Mencius* (I. Bloom, Trans.). New York, NY: Columbia University Press.

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

Newman, G. E., Lockhart, K. L., & Keil, F. C. (2010). "End-of-life" biases in moral evaluations of others. *Cognition*, *115*, 343-349.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*, 586-604.

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

Pinel, E. C., Long, A. E., Laundau, M., Stanley, K., & Pyszczynski, T. (2006). Seeing I to I: A pathway to interpersonal connectedness. *Journal of Personality and Social Psychology*, *90*, 243-257.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279-301.

Rozenblit, L. R., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

Schimel, J., Arndt, J., Pyszczynski, T., & Greenberg, J. (2001). Being accepted for who we are: Evidenced that social validation of the intrinsic self reduces general defensiveness. *Journal of Personality and Social Psychology*, *80*, 35-52.

Schlegel, R. J., Hicks, J. A., Arndt, J., & King, L. A. (2009). Thine own self: True self-concept accessibility and meaning in life. *Journal of Personality and Social Psychology*, *96*, 473-490.

Schlegel, R. J., Hicks, J. A., Davis, W. E., Hirsch, K. A., & Smith, C. M. (2013). The dynamic interplay between perceived true self-knowledge and decision satisfaction. *Journal of Personality and Social Psychology*, *104*, 542-558.

Schlegel, R. J., Hicks, J. A., King, L. A., & Arndt, J. (2011). Feeling like you know who you are: Perceived true self-knowledge and meaning in life. *Personality and Social Psychology Bulletin*, *37*, 745-756.

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*, 159-176.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance-estimation bias. *Psychonomic Bulletin and Review*, *4*, 387-392.