COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2014) 1–30

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12134

Beliefs About the True Self Explain Asymmetries Based on Moral Judgment

George E. Newman,[a] Julian De Freitas,[b] Joshua Knobe[c]

[a]*Yale School of Management, Yale University*
[b]*Department of Psychology, Yale University*
[c]*Program in Cognitive Science and Department of Philosophy, Yale University*

Received 30 May 2013; received in revised form 27 September 2013; accepted 15 October 2013

Abstract

Past research has identified a number of asymmetries based on moral judgments. Beliefs about (a) what a person values, (b) whether a person is happy, (c) whether a person has shown weakness of will, and (d) whether a person deserves praise or blame seem to depend critically on whether participants themselves find the agent's behavior to be morally good or bad. To date, however, the origins of these asymmetries remain unknown. The present studies examine whether beliefs about an agent's "true self" explain these observed asymmetries based on moral judgment. Using the identical materials from previous studies in this area, a series of five experiments indicate that 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 morally virtuous. In turn, this belief causes people to hold different intuitions about what the agent values, whether the agent is happy, whether he or she has shown weakness of will, and whether he or she deserves praise or blame. These results not only help to answer important questions about how people attribute various mental states to others; they also contribute to important theoretical debates regarding how moral values may shape our beliefs about phenomena that, on the surface, appear to be decidedly non-moral in nature.

Keywords: Concepts; Social cognition; Moral reasoning; True self; Happiness; Values; Weakness of will; Blame

1. Introduction

Recent research has led to a surprising convergence on what might at first seem to be unrelated topics. In particular, past work has examined people's intuitions about *valuing*

Correspondence should be sent to George E. Newman, Organizational Behavior and Cognitive Science, 165 Whitney Ave., Yale University, New Haven, CT 06520. E-mail: george.newman@yale.edu Organizational Behavior and Cognitive Science165 Whitney Ave.06520

(Knobe & Roedder, 2009), *happiness* (Phillips, Misenheimer, & Knobe, 2011; Phillips, Nyholm, & Liao, in press), *weakness of will* (May & Holton, 2012), and *moral responsibility* (Pizarro, Uhlmann, & Salovey, 2003). Although research in these areas has proceeded separately, these different lines of work have arrived at the same basic finding: In each case, researchers have observed an asymmetry between the intuitions people have about behaviors that are morally good versus those that are morally bad.

We will be discussing each of the relevant effects in further detail below, but very briefly, the pattern of the four asymmetries is as follows:

1. The *valuing asymmetry* arises in cases where an agent's desires conflict with her beliefs. If the agent ends up acting on her desires, participants are more inclined to say that she "values" what she is doing when they believe that her actions are morally good compared to when they are morally bad (Knobe & Roedder, 2009).
2. The *happiness asymmetry* arises in cases where an agent has positive emotions and is satisfied with her life. In such cases, participants are more inclined to say that the agent is "happy" when they believe that her actions are morally good compared to when they are morally bad (Phillips et al., 2011, in press).
3. The *weakness of will asymmetry* arises in cases where an agent experiences a conflict between her desires and her beliefs, and ultimately acts on her desire (thereby going against her belief). In such cases, participants are less inclined to say that the agent shows "weakness of will" if they regard the desire as morally good than if they regard the desire as morally bad (May & Holton, 2012).
4. The *blame/praise asymmetry* arises in cases where an agent is overwhelmed by emotion and cannot stop herself from performing an action. In such cases, participants are inclined to say that she deserves decreased blame when they believe that her action is morally bad, but she does not deserve decreased praise when they believe that her action is morally good (Pizarro et al., 2003).

These asymmetries are puzzling because many of them arise even for responses to questions that one might expect to be entirely independent of participants' own moral judgments. For example, when participants are trying to determine whether an agent "values" a particular outcome, one might expect them to focus on facts about the agent's psychological states (i.e., the agent's beliefs, desires, emotions, etc.)—it doesn't appear that they would have to assess the moral value of those states (e.g., deciding whether the agent's desires were morally good or morally bad). Yet the experimental results indicate that participants are doing precisely that. In each case, researchers provide the same basic information about the agent's psychological states, manipulating only information that would be relevant to moral judgments, and in each case, they find an effect. The question we explore here is, why are all of these effects arising, and can they be explained under a single psychological framework?

In the existing literature, each of the asymmetries is regarded as a difficult problem that would be worthy of further research. (For a few discussions, see Beebe, 2013; Gonnerman, 2008; Knobe & Doris, 2010). However, the usual tendency is to examine each of these asymmetries in isolation from all the others and to seek separate explanations for

each of the asymmetries—one for the valuing asymmetry, one for the weakness of will asymmetry, and so on. In fact, as far as we can tell, there have not yet been any articles that mention all four asymmetries.

Our aim here was to pursue the opposite approach. We propose that these four asymmetries are actually best understood as four different symptoms of the same underlying phenomenon. In particular, we will argue that all four can actually be explained in terms of the way that people's moral judgments impact their understanding of the *true self*.

1.1. The concept of a "true self"

The concept of a "true self" is the concept people employ when they speak of "being true to yourself" or "revealing the person you really are, deep down inside." Though intellectuals of various stripes have claimed that this whole notion is a mistaken or incoherent one (Foucault, 1984; Sartre, 1958/2003), empirical research consistently finds that people's ordinary understanding of the mind does involve a distinction between a "true self" (sometimes referred to as a person's "core" or "essence") and more superficial aspects of the self (sometimes known collectively as the "false self"; Johnson, Robinson, & Mitchell, 2004).

Existing work has shown that people's judgments about which aspects of a person's mind fall within the true self have substantial impacts on a number of other psychological processes. Such judgments have been shown to influence moral 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 (Baumeister, 1991; Schlegel, Hicks, Davis, Hirsch, & Smith, 2013), as well as more general measures of well-being (e.g., Kernis & Goldman, 2004, 2006; Schimel, Arndt, Banko, & Cook, 2004; Schimel, Arndt, Pyszczynski, & Greenberg, 2001). The key question, however, is how people distinguish between the aspects of a person's mind that fall inside the true self and those that fall outside it.

Initially, it might appear that this distinction can be reduced to more familiar ones, such as the distinction between reason and emotion. Within the philosophical literature, for example, it has long been claimed that an agent's true self can be identified with the more reflective aspects of that agent's mind (Aristotle, 350 BC/1985; Frankfurt, 1971). On such a view, if an agent reflects carefully and determines that one of her desires is deeply wrong, then the right thing to conclude is that this more reflective conclusion is revealing her true self and that the desire is not part of the true self at all. Conversely, 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 reflection (e.g., Gide, 1902). On this latter view, the actions that most fully reveal an agent's true self are the ones she performs when she is so overcome with emotion that she cannot control herself. Although these two views are exact opposites of each other, they both make the notion of a true self seem relatively straightforward. All one has to do to determine whether an action

reflects an agent's true self is to figure out which aspect of the mind (e.g., reason vs. emotion) this action stemmed from.

Existing studies suggest, however, that people's ordinary understanding of the true self does not conform to either of these simple views (Newman, Bloom, & Knobe, 2014). People do not appear to consistently identify the true self either with reason or with emotion. Instead, people's true self attributions appear to be influenced in a complex way by their *moral* judgments. That is, when people are trying to determine whether a given psychological state falls within the agent's true self, they seem to be influenced by their beliefs about whether this state itself is morally good or morally bad.

In one study illustrating this effect (Newman et al., 2014), participants were given a vignette about an agent who was experiencing an inner conflict. Participants in the "pro-gay belief" condition were told that the agent experienced a strong emotional distaste toward anyone who engaged in homosexual behavior but that, when he thought the matter over in a more reflective way, he concluded that homosexuality was perfectly acceptable. Participants in the "anti-gay belief" condition then received a vignette with precisely the opposite structure: The agent had a strong desire to sleep with other men, but when he thought the matter over in a more reflective way, he concluded that homosexuality was morally wrong. Participants in both conditions were asked whether the agent's belief was part of his true self.

Both of the simple theories would say that the answer to this question should be straightforward. According to the view that an agent's more reflective beliefs constitute the true self, the belief would be part of the true self in both conditions. Conversely, according to the view that an agent's more unreflective emotions and cravings are the true self, the belief would not be the true self in either case. What the results showed, however, was that participants were not drawn toward either of these simple views. Instead, their beliefs about the agent's true self reflected their own value judgments: Politically liberal participants tended to say that the pro-gay belief was part of the agent's true self but the anti-gay belief was not, while politically conservative participants tended to say that the anti-gay belief was part of the agent's true self but the pro-gay belief was not (Newman et al., 2014; Study 3). This result provides initial evidence for the view that people's judgments as to whether a belief falls within the true self might depend in part on their judgments about whether that belief is morally right or morally wrong (see Newman et al., 2014 for further examples).

Drawing on this result, we propose a general hypothesis about people's true self attributions. Such attributions may be affected by numerous different factors, but all else being equal, we propose that people are inclined to attribute to the true self mental states that *they themselves regard as morally good*. Thus, suppose that an agent is described as having certain beliefs, desires, and emotions. Independent of the differences between these different types of states, people will be inclined to see whichever ones they regard as morally good as lying within the true self and whichever ones they regard as morally bad as lying outside of it.

Note that this hypothesis has clear implications for the studies that originally demonstrated the four asymmetries. In each of these studies, all participants received information

about an agent's psychological states (e.g., that the agent was experiencing a conflict between beliefs and emotions). Researchers then systematically varied the moral status of these states (e.g., a morally good belief in one condition vs. a morally bad belief in the other). It might seem at first that the only thing being varied in these studies is the information about moral status—information about the agent's actual psychological states is remaining constant across conditions. According to the present hypothesis, however, this is not the case. Instead, the changes in moral status are leading immediately to changes in participants' conception of the agent's psychological states. In particular, when a state is presented in a way that makes it seem morally good, participants tend to see it as falling inside the agent's true self, whereas when a state is presented in a way that makes it seem morally bad, participants tend to see it as falling outside the agent's true self.

This basic framework now opens the door to a new explanation for the asymmetries. Specifically, our proposal is that all of the asymmetries are arising because people's moral judgments have an impact on their true self attributions, and these true self attributions in turn impact their intuitions about valuing, happiness, weakness of will, and moral responsibility.

1.2. The present studies

To test this hypothesis, we use a method that was first introduced in work on intuitions about intentional action (Sripada & Konrath, 2011). Earlier studies had pointed to an asymmetry in people's intuitions about whether an agent could be said to be acting "intentionally" (Knobe, 2003), and Sripada and Konrath (2011) hypothesized that this asymmetry might be best explained in terms of people's understanding of that agent's self. To test this, they replicated an earlier study while adding a new item that measured judgments about the agent's self, and showed that the impact of condition on attributions of intentional action was mediated by judgments about the self (Sripada & Konrath, 2011). We will be using that same methodology here. (In the General Discussion, we return to the intentional action asymmetry and ask whether it might be related to the four asymmetries explored in the present studies.)

As Fig. 1 shows, the hypothesis under consideration here predicts a complex pattern of mediation. Specifically, the hypothesis posits two distinct steps: (a) moral judgments impact true self attributions and then (b) true self attributions impact the use of each of the concepts for which the asymmetries have been shown (valuing, happiness, weakness of will, moral responsibility). Thus, the hypothesis predicts that true self attributions should mediate the impact of moral judgments on the application of each of these concepts. To test this hypothesis, Experiments 1–4 used bootstrap mediation (Preacher & Hayes, 2008). In each case, we replicated existing work demonstrating the relevant asymmetry, and then showed that this asymmetry was mediated by true self attributions. In a final study (Experiment 5), we manipulated beliefs about the true self directly and measured the resulting effects on the application of each of the concepts.

One persistent worry is that when researchers have sufficient degrees of freedom, they can find evidence for just about anything they want (see Simmons, Nelson, & Simonsohn,

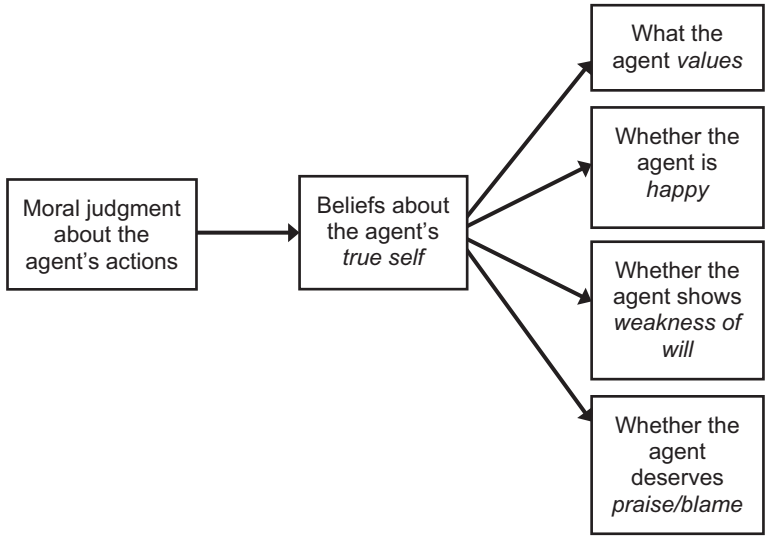

Fig. 1. Hypothesized mediation pattern, with the impact of moral status on each of the four concepts (valuing, happiness, weakness of will, and praise/blame) mediated by true self attributions.

2011 for the general worry; Strickland & Suben, 2012 for vignette studies in particular). To address this worry, we conducted these studies in a way designed to drastically restrict our own degrees of freedom. First, we examined each of the four asymmetries using the same basic experimental design and the same statistical analyses. Second, we examined each of these asymmetries using stimulus materials that had already been developed within existing research.

This last restriction leaves us with an especially strong test of the hypothesis. The present experiments all made use of stimulus materials that were originally constructed by researchers who were not in any way trying to study true self attributions. Nonetheless, we predict that each of these stimuli will show an impact of moral valence on true self attributions and, more important, that this impact of valence on true self attributions will mediate the previously shown asymmetries.

Much of the theoretical work needed to motivate our hypothesis will involve discussing each of the four asymmetries separately, and we will be discussing those theoretical questions in the introductory sections of each of the separate experiments. But the experiments as a whole aim to do more than simply address four separate issues: Taken together, they point to the true self as a core explanatory concept in people's attributions about others.

2. Experiment 1: The valuing asymmetry

One of the most central concepts within philosophical work on moral agency is the concept of *valuing* (Lewis, 1989; Smith, 1994; Watson, 1975). Research using this concept has emphasized the difference between merely having a desire and actually having a value. For

a particularly clear example, imagine a frustrated parent who wants to treat her child fairly but who sometimes gets upset and feels an urge to start screaming. In a case like this, it might be correct to attribute to the parent two opposing desires: "She wants to treat her child fairly, but sometimes she also wants to start screaming." However, only one of these desires would count as a "value." Thus, it might be right to say, "She values treating her child fairly," but it would be wrong to say, "She values screaming at her child" (see Lewis, 1989; Smith, 1994; Watson, 1975). The key question is then how to understand this concept of valuing and what it involves that goes beyond the concept of desire.

Philosophical work on this issue has tended to focus on the hypothesis that for an agent to truly value something, she has to approve of it on a more reflective level. This hypothesis has been spelled out in a number of different ways. For example, Smith (1994) suggests that an agent cannot be said to value something unless she believes that it is good, while Lewis (1989) argues that an agent cannot be said to value something unless she has a second-order desire for it. Either way, the basic idea is clear enough. The parent in our example can be said to value fairness (but not screaming) because she approves of fairness on a more reflective level.

A series of experimental studies, however, have provided evidence that these philosophical theories do not fit the patterns of people's ordinary judgments (Knobe & Roedder, 2009). In one study, participants were given a story about an agent who is caught between a desire for racial equality and a desire for racial discrimination. Participants in one condition were told that the agent firmly believes in racial equality but nonetheless has an urge to act in ways that promote racial discrimination:

> Jim lives in a culture in which most people believe in racial equality. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of all people equally, regardless of their race.
>
> Nonetheless, Jim sometimes feels a pull in the opposite direction and sometimes he ends up acting on these feelings and doing things that end up fostering racial discrimination.
>
> Jim wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial discrimination and just act to advance the interests of all people equally, regardless of their race.

Participants in the other condition received the converse case (loosely based on Twain, 1885/1994, character Huck Finn), involving an agent who firmly believes in racial discrimination but nonetheless has an urge to act in ways that promote racial equality:

> Jim lives in a culture in which most people are extremely racist. He thinks that the basic viewpoint of people in this culture is more or less correct. That is, he believes that he ought to be advancing the interests of people of his own race at the expense of people of other races.

Nonetheless, Jim sometimes feels a pull in the opposite direction and sometimes he ends up acting on these feelings and doing things that end up fostering racial equality.

Jim wishes he could change this aspect of himself. He wishes that he could stop feeling the pull of racial equality and just act to advance the interests of his own race.

Existing philosophical theories would say that, as the agent in both of these cases does not approve of his urge on a more reflective level, neither of the urges are examples of valuing. Yet people's ordinary judgments show an asymmetry: They tend to say that the agent in the first case did not value racial discrimination but that the agent in the second case did value racial equality (Knobe & Roedder, 2009). In other words, people's concept of valuing seems to be more closely tied to their own moral values (e.g., the belief that equality is good, while discrimination is bad), rather than the extent to which, upon reflection, the agent in the vignette endorses the particular mental state.

A number of different explanations have been proposed for this effect (Gonnerman, 2008; Kauppinen, 2006; Knobe & Roedder, 2009), but none of these explanations has gained widespread support, and the issue is widely regarded as unresolved. We propose that the effect can be explained using the theoretical framework introduced here.

The key hypothesis is that people regard a psychological state as "valuing" to the extent that it falls within the agent's *true self* (which itself is assumed to be fundamentally good). In cases where the agent's more reflective states draw him toward actions that are morally good, people will tend to identify the reflective state as in line with the agent's true self (and their judgments will therefore fit the traditional philosophical analyses where the reflective state is seen as one of the agent's values). By contrast, in cases where the agent's more reflective states draw him toward actions that are morally bad, people will see these states as lying outside the agent's true self (and their judgments will go against the traditional analyses—they will conclude that the reflective mental state is not one of the agent's values). Thus, the asymmetry observed for valuing judgments simply falls out of the asymmetry observed more generally for judgments about the true self.

To test this hypothesis, we replicated an existing study on judgments of valuing (Knobe & Roedder, 2009). The materials were identical to the previous study, but we added an item about the true self. The key question was whether the impact of condition on valuing judgments would be mediated by this true self judgment. We also asked participants about the agent's feelings and beliefs, to contrast these answers against ratings of the true self. For example, perhaps the asymmetries in "valuing" judgments emerge because participants doubt that the agent believes in racial discrimination to the same extent that he believes in racial equality.

2.1. Method

Participants were 54 undergraduates (M_{age} = 19.4, 59% female) who were recruited on campus to participate in a psychological study in exchange for $3.

Participants received the same vignettes as used in the Knobe and Roedder (2009) study described above, involving an agent whose feelings drive him either toward racial discrimination or toward racial equality. In the racial discrimination condition, participants were asked to rate the extent to which they agreed with the statement, "Despite his conscious beliefs, Jim actually values racial discrimination." In the racial equality condition, participants were asked to rate the extent to which they agreed with the statement, "Despite his conscious beliefs, Jim actually values racial equality." Both statements were rated on a scale from 1 ("strongly agree") to 9 ("strongly disagree").

Following the valuing measure, participants rated their agreement with the following three items using the same scale as the valuing measure (1 = *strongly agree*, 9 = *strongly disagree*): "Jim was being drawn toward racial discrimination by his feelings." "Jim was being drawn toward racial discrimination by his beliefs." "Jim was being drawn toward racial discrimination by his true self—the person he really is deep down." In the condition that asked about valuing racial equality, all aspects of the study were identical except that the phrase "racial discrimination" was replaced with the phrase "racial equality."

2.2. Results

We first examined differences between the two conditions (racial equality vs. racial discrimination) across the four measures. (Means for all measures in Experiments 1–4 are reported in Table 1. Note that, in this study, lower numbers indicate higher agreement.)

Table 1
Means for all measures in Experiments 1–4

Experiment 1	True Self	Values	Beliefs	Feelings
Good	4.07	4.41	6.33	3.93
Bad	5.63	6.63	6.59	4.22
p	<.01	<.001	.63	.62
Experiment 2	True Self	Happiness	Moral Judgment	
Good	3.93	6.25	5.78	
Bad	5.35	5.21	1.59	
p	<.001	<.001	<.001	
Experiment 3	True Self	Weakness of Will	Decision	
Good	2.67	2.74	5.90	
Bad	3.42	5.73	6.36	
p	.02	<.001	.04	
Experiment 4	True Self	Praise/Blame	Meta-desires	
Bad				
Impulsive	6.11	7.24	4.81	
Deliberate	7.11	7.93	6.18	
Good				
Impulsive	7.14	6.66	5.53	
Deliberate	7.18	6.96	6.76	
p (interaction)	.02	.08	.72	

Consistent with past research (Knobe & Roedder, 2009), we observed that participants were significantly more likely to agree that the agent valued racial equality (M = 4.41, SD = 2.21) than racial discrimination (M = 6.63, SD = 1.74), $t(52) = 4.11$, $p < .001$. In addition, participants were significantly more likely to agree that the agent's true self was drawn toward racial equality (M = 4.07, SD = 1.94) than toward racial discrimination (M = 5.63, SD = 2.01), $t(52) = 2.83$, $p = .007$. In contrast, ratings of the agent's feelings (Ms = 3.93, 4.22) and the agent's beliefs (Ms = 6.33, 6.59) did not significantly differ across the two conditions, ps = .62 and .63, respectively. However, agreement with the feelings measure (M = 4.07) was overall significantly greater than the beliefs measure (M = 6.46), which was consistent with the factual information presented in each vignette, $F(1,52) = 32.52$, $p < .001$.

To test the mediating role of beliefs about the true self, we then conducted a bootstrap multiple-mediators analysis (Preacher & Hayes, 2008) with condition as the independent variable (dummy coded, 1 = *racial equality*, 0 = *racial discrimination*), ratings of "valuing" as the dependent variable, and measures of the true self, feelings, and beliefs as potential mediators. This analysis indicated that only ratings of the true self significantly mediated the effect of condition on the measure of valuing (95% CI = −1.64 to −.12; see Fig. 2).

2.3. Discussion

Results from this first study were consistent with our hypotheses. Replicating previous research (Knobe & Roedder, 2009), we found that in cases where an agent did not approve of his urge on a more reflective level, participants were nonetheless more likely to agree that the urge counted as one of the agent's values when it was morally good (racial equality) compared to when it was morally bad (racial discrimination). Moreover, we found via mediation analyses that belief in the true self explained this effect such that the critical difference across conditions was the extent to which participants viewed that

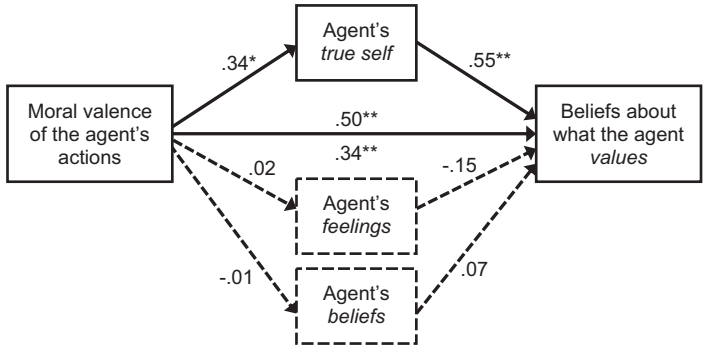

Fig. 2. Mediation results from Experiment 1. *$p < .05$; **$p < .001$.

urge as reflective of the agent's true self. By contrast, an alternative explanation based on whether the good behavior was more reflective of the agent's beliefs than desires did not appear to explain the effect.

3. Experiment 2: The happiness asymmetry

Our next asymmetry arises in people's use of the concept *happiness*. This concept figures not only in people's ordinary judgments but also in scientific research, and work in this more scientific tradition has led to a substantial degree of convergence about how to understand the nature of happiness. The consensus view is that happiness involves a high level of positive emotion, a low level of negative emotion, and a high degree of life satisfaction (for a review, see Diener, Scollon, & Lucas, 2003).

But here again, a series of studies indicate that people's ordinary judgments show a moral asymmetry (Phillips et al., 2011, in press). In one study, all participants received vignettes about an agent who had the three psychological features emphasized by existing theoretical accounts (high positive emotion, low negative emotion, high satisfaction). The only difference between conditions was in the information about how the agent was living her life. In one condition, the agent was described as living a morally good life. For example:

> Maria is the mother of three children who all really love her. In fact, they couldn't imagine having a better mom. Maria usually stays pretty busy taking care of her children. She often finds herself rushing from one birthday party to the next, visiting friends important to her children.
>
> Almost every single day Maria feels good and generally experiences a lot of pleasant emotions. In fact, it is very rare that she would ever feel negative emotions like sadness or loneliness. When Maria thinks about her life, she always comes to the same conclusion: she feels highly satisfied with the way she lives.
>
> The reason Maria feels this way is that while Maria has been preoccupied with her children, she still makes an effort to be nice and spend time with her old friends. Almost every night she ends up helping her children or planning something for her children's future.

In the other condition, the agent was described as living a morally bad life. For example:

> Maria wants to live the life of a celebrity in L.A. In fact, she has even started trying to date a few famous people. Maria usually works hard to become popular. She often finds herself rushing from one social gathering to the next and is always going to pick up some alcohol or a dress.

> Almost every single day Maria feels good and generally experiences a lot of pleasant emotions. In fact, it is very rare that she would ever feel negative emotions like sadness or loneliness. When Maria thinks about her life, she always comes to the same conclusion: she feels highly satisfied with the way she lives.
>
> The reason Maria feels this way is that Maria is so preoccupied with becoming popular that she is no longer concerned with being honest or nice to her old friends unless they know someone famous. Almost every night she ends up drinking or partying with famous people she wants to be like.

Given these descriptions of the agent's psychological states, existing theoretical accounts would say that she is clearly happy in both of these vignettes. However, experimental participants show an asymmetry. They report that the agent in the morally good case is happy, but they show a marked reluctance to say that the agent in the morally bad case is happy (Phillips et al., 2011).

We suggest that this asymmetry, too, can be explained within the present theoretical framework. The key aspect of the theory is the claim that even when all available evidence suggests that an agent's psychological states are drawing her to behave in ways that are morally bad, people are still inclined to posit a hidden "true self" that is calling her to behave in ways that are morally good. In the studies under discussion here, participants in both conditions are told about an agent who appears on the surface to be entirely satisfied with her life. However, if the theory we have been developing is correct, participants might be inclined in some cases to posit a deeper level at which the agent actually feels quite different from the way she appears to feel on the surface.

More specifically, participants should show an asymmetric pattern of judgments. In the case where the agent has a morally good life and appears to feel satisfied with the way she has been living, there is no reason to expect participants to posit a deeper level on which she actually feels dissatisfied. By contrast, in the case where the agent has a morally bad life, participants should be inclined to posit a deeper level on which her "true self" is calling her to lead a morally good life. So even if they recognize that she is happy at a superficial level, they should think that there is a deeper level on which she is fundamentally dissatisfied with the way she has been living.

To test this hypothesis, we used the same method as in the previous experiment: replicating an existing study, adding an item about the true self, and testing for mediation.

3.1. Method

Participants were 161 adults (M_{age} = 28.9, 37% female) who were recruited from mTurk and participated in exchange for \$0.25. Participants were assigned to one of eight conditions in a 2 (moral valence: good vs. bad) × 4 (vignette) design. The difference across vignettes was not hypothesized to be an important factor and served merely as a robustness check.

Participants read one of eight vignettes about an agent who had three psychological features typically associated with happiness: high positive emotion, low negative emotion, and high satisfaction. Half of the participants read about an agent who engaged in morally good behaviors, while the other half read about an agent who engaged in morally bad behaviors. The vignettes were very similar in structure and consisted of the following matched pairs: Woman who cares for her children versus woman who only likes to party; nurse who helps patients versus nurse who poisons patients; janitor who assists disabled students versus janitor who steals from students; man who cares for his niece versus man who molests his niece.

Following each vignette, participants then rated their agreement with the following statement: "[Maria] is happy" (1 = *strongly disagree*, 7 = *strongly agree*); the agent's name was changed to match the particular vignette. On a subsequent page, all participants rated their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with two additional statements, one that assessed beliefs about the true self, and one that assessed participants' own moral judgments about the behavior. Specifically, participants responded to the following: "Deep down, Maria actually feels very differently about her life from the way she feels on the surface" and "Maria is living the kind of life she should be living." The order in which each of these items appeared was counterbalanced across participants.

3.2. Results

A 2 (valence: good vs. bad) × 4 (vignette) ANOVA indicated a significant main effect of valence on ratings of happiness, $F(1,153) = 24.84$, $p < .001$. Replicating Phillips et al. (2011), we observed that participants were significantly more likely to agree that the agent was happy when the behavior was morally good ($M = 6.25$, $SD = 0.82$) than when the behavior was morally bad ($M = 5.21$, $SD = 1.78$). There was also a main effect of vignette, where participants gave higher happiness ratings to some vignettes over others, $F(3,153) = 6.35$, $p < .001$, though importantly, this factor did not interact with the primary variable of moral valence, $p = .275$.

We then conducted a mediation analysis to determine whether beliefs about the true self explained the effect of moral valence on ratings of happiness. For this particular study, we were able to conduct a serial mediation model (Preacher & Hayes, 2008), with the prediction that participants' own moral judgments about the agent's behavior should in turn influence their beliefs about the true self, which should subsequently influence ratings of happiness (i.e., target's behavior → moral judgment of the behavior → beliefs about the true self → ratings of happiness). The bootstrap analysis indicated that this serial mediation model was indeed significant using the predicted pathway of moral judgments of the behavior → beliefs about the true self (95% CI = .17–.95; see Fig. 3). In contrast, the reverse mediation model (target's behavior → beliefs about the true self → moral judgment of the behavior → ratings of happiness) was not significant (95% CI = −.01 to .05).

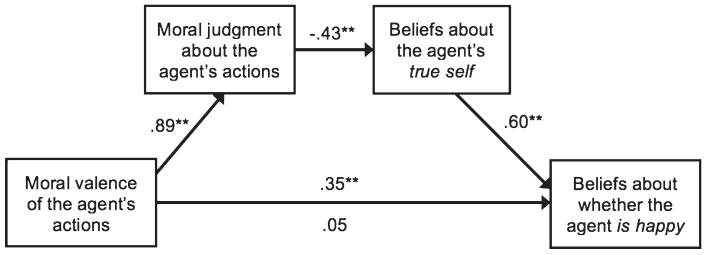

Fig. 3. Mediation results from Experiment 2. $^{**}p < .001$.

3.3. Discussion

Results from this second study lend further support to the notion that the true self plays a key role in explaining asymmetries based on moral judgments. Although all the agents in these vignettes exhibited the "classic" features of happiness (a high level of positive emotion, a low level of negative emotion, and a high degree of life satisfaction; Diener et al., 2003), those who lived a morally good life were judged to be happier than those who lived a morally bad one (Phillips et al., 2011). Moreover, we identified the specific mediation pathway underlying this effect. As hypothesized, participants judged the morally good vignettes to be normatively better than the morally bad vignettes. In turn, this difference gave rise to different beliefs about the true self—namely, the morally bad behaviors conflicted with the agent's true self, while the morally good behaviors did not. In other words, in the case where the agent lived a morally good life, participants did not posit a deeper level on which she actually felt dissatisfied, but in the case where the agent lived a morally bad life, participants did posit a deeper level of dissatisfaction. Finally, the mediation results indicated that it was the difference in true self beliefs that ultimately explained the asymmetry in participants' judgments about the agent's happiness.

4. Experiment 3: The weakness of will asymmetry

Consider the following brief vignette (taken from Sousa & Mauro, 2013):

John is a professional assassin. He has started to think about quitting this profession because he feels that it is wrong to kill another person. However, he is strongly inclined to continue with it because of the financial benefits.

John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to quit his profession. Accordingly, he decides that the next day he will look for a job that does not involve violence.

> The next day, while still completely sure that the best thing for him to do is to look for a job that does not involve violence, John is swayed by the financial benefits. Against what he had decided, he kills another person for money.

This is a paradigm case of what has traditionally been called *weakness of will* (Aristotle, 350 BC/1985). In some sense, it seems that the man described here has given in to temptation and is failing to exercise control over his own self.

Within the philosophical literature, the usual approach to analyzing the notion of weakness of will starts out with the idea that a weak-willed agent performs actions that go against what he or she intends or believes to be right (Aristotle, 350 BC/1985). Difficult questions arise about precisely how to spell out this basic approach, and different philosophical analyses differ in important respects (Holton, 1999; Mele, 2010). Still, despite these differences, most existing philosophical accounts would agree, at least in broad outline, about what makes the action in the story above count as an example of weakness of will. They would say that the action counts as weak-willed because (a) the agent intended to quit his profession and believed that it would be best to do so but then (b) the agent actually did just the opposite, continuing to do precisely the things that his intentions and beliefs pointed against.

However, a series of recent experimental studies show that people's ordinary judgments about weakness of will do not fit these traditional philosophical analyses. Instead, people's judgments about weakness of will appear to be influenced by the *moral* status of the action that the agent performs (May & Holton, 2012; replicated and extended in Beebe, 2013; see also Doucet & Turri, 2013; Sousa & Mauro, 2013). Even controlling for all of the factors that philosophers have traditionally seen as relevant, people appear to be more inclined to say that an action displays weakness of will when they themselves believe that the agent's action is morally wrong.

In a particularly elegant demonstration of this effect (Sousa & Mauro, 2013), participants in one condition received precisely the vignette about the assassin quoted above (where the agent's action is clearly morally wrong), while participants in the other condition received a modified version in which the agent's action is morally right:

> John is in conflict, but after considering all aspects of the matter, he concludes that the best thing for him to do is to continue with his profession. Accordingly, he decides that the next day he will kill another person for money.
>
> The next day, while still completely sure that the best thing for him to do is to kill another person for money, John is swayed by the feeling that it is wrong to kill. Against what he had decided, he looks for a job that does not involve violence.

Note that in both conditions, the assassin goes against what he intends and believes to be best, so existing philosophical analyses would say that the agent showed weakness of will in both conditions. Yet this was not at all how experimental participants actually reacted. Instead, participants tended to say that the agent showed weakness of will in the

condition where his action was morally wrong, but not in the condition where his action was morally right (Sousa & Mauro, 2013).

This basic effect has now emerged in a number of different studies from different laboratories (Beebe, 2013; May & Holton, 2012; Sousa & Mauro, 2013), but there is no agreement on how the effect is to be explained; the issue is widely regarded as an open question. As in the previous studies, our proposal is that the best way to explain these phenomena might be in terms of people's conception of the true self.

The first step in this explanation is to modify the usual view about the concept of weakness of will. We argue against the hypothesis that people have a general tendency to regard an action as weak-willed to the extent that it goes against an agent's intentions or beliefs. Instead, we suggest that people regard an action as weak-willed to the extent that it is seen as going against an agent's true self. In cases where the agent's belief is seen as morally right, this belief should be seen as part of the true self. However, in cases where his belief is seen as morally wrong, the agent should be regarded very differently: The belief should be seen as lying outside the true self, such that when his actions go against this belief, they are not judged as weak-willed. In short, our hypothesis is that the impact of moral considerations on judgments of weakness of will is best understood as simply falling out of the impact of moral considerations on judgments about the true self.

To test this hypothesis, we used the same approach as the previous experiments: replicating an existing study on intuitions about weakness of will, but adding an item assessing beliefs about the true self. The key question was whether judgments about the true self would mediate the impact of condition on judgments about weakness of will.

4.1. Method

Participants were 139 adults (M_{age} = 37.7, 46% female) who were recruited from mTurk and participated in exchange for $.25. Participants were assigned to one of four conditions in a 2(moral valence: good vs. bad) × 2(vignette) design. Again, the difference across vignettes was not hypothesized to be an important factor and served merely as a robustness check. All materials (except for the added mediation measures) were identical to those used in Sousa and Mauro (2013).

Participants read one of four vignettes about an agent who had an intention to do one thing but wound up doing the exact opposite. For half of the participants, the intention was morally good (e.g., not being an assassin), while for the other half the intention was morally bad (e.g., continuing to be an assassin). The specific vignettes were the assassin example presented above and a nearly identical "robber" example that involved stealing from rather than killing people.

Following each vignette, participants then rated their agreement with the following statement (1 = *strongly disagree*, 9 = *strongly agree*): "John displays weakness of will when, the next day, he kills (steals from) another person for money." On a subsequent page, all participants rated their agreement (1 = *strongly disagree*, 9 = *strongly agree*) with two additional statements, one that asked about the agent's decision and one that asked about the agent's true self. Specifically, participants responded to the following:

"What John did at the end of the story went against the decision he made earlier" and "What John did at the end of the story went against his true self—the person he truly is deep down." The order in which each of these items appeared was counterbalanced across participants.

4.2. Results

A 2 (valence: good vs. bad) × 2 (vignette) ANOVA indicated a significant main effect of valence on ratings of "weakness of will," $F(1,135) = 103.24$, $p < .001$. Replicating Sousa and Mauro (2013), we observed that participants were significantly less likely to agree that the agent showed weakness of will when the behavior was morally good ($M = 2.74$, $SD = 1.90$) than when the behavior was morally bad ($M = 5.73$, $SD = 1.52$). There was no main effect of vignette, $p = .84$, and no interaction, $p = .33$.

To test the mediating role of beliefs about the true self, we then conducted a bootstrap multiple-mediators analysis (Preacher & Hayes, 2008) with condition as the independent variable (dummy coded, 1 = *bad action*, 0 = *good action*), and ratings of "weakness of will" as the dependent variable. The potential mediators were "Did the agent go against his true self?" (true self) and "Did the agent go against his decision?" (decision). This analysis indicated that ratings of the true self significantly mediated the effect of condition on perceptions of weakness of will (95% CI = .03–.44; see Fig. 4). There was also a marginal indirect effect of decision on weakness of will (95% CI = .009–.29).

4.3. Discussion

This study replicated the "weakness of will" effect whereby participants tended to say that the agent showed weakness of will in the condition where his action was morally wrong, but not in the condition where his action was morally right (May & Holton, 2012; Sousa & Mauro, 2013). Consistent with our predictions (and the results of the previous

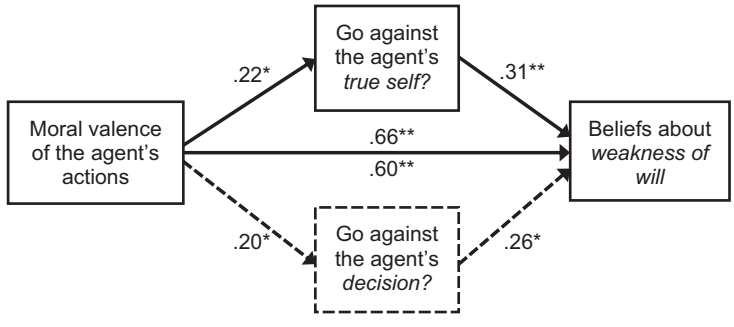

Fig. 4. Mediation results from Experiment 3. *$p < .01$; **$p < .001$.

studies reported here) expectations about the true self mediated this effect. Specifically, we found that people regarded an action as weak-willed to the extent that it was seen as going against the agent's true self—in cases where the agent's initial decision was morally right, the action was seen as part of the true self more so than when the initial decision was morally wrong. In turn, this asymmetry in beliefs about the true self led participants to have different beliefs about the extent to which the agent showed weakness of will.

5. Experiment 4: Blame and praise for impulsive behavior

The final asymmetry arises in people's assignment of moral responsibility. Specifically, Pizarro et al. (2003) found a striking asymmetry in how people assign praise and blame in cases where an agent acts impulsively. For example, consider the following pair:

> Deliberate: Jack calmly and deliberately smashed the window of the car parked in front of him, because it was parked too close to his.
>
> Impulsive: Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him, because it was parked too close to his.

In this case, participants made an important distinction between deliberate and impulsive behaviors, assigning significantly more blame when the behavior was deliberate than when it was impulsive. However, when the behavior was morally good, participants exhibited a very different pattern of responses. For example:

> Deliberate: Jack deliberately and intentionally gave the homeless man his only jacket, even though it was freezing outside.
>
> Impulsive: Because of his overwhelming and uncontrollable sympathy, Jack impulsively gave the homeless man his only jacket, even though it was freezing outside.

In this case, participants did not assign different levels of praise to the actions that were performed deliberately versus impulsively.

Thus, there is an asymmetry in how impulsive actions are evaluated in the context of either morally good or morally bad behavior (Pizarro et al., 2003). If an agent is so overcome with emotion that he cannot prevent himself from doing something morally bad, people tend to say that he deserves reduced blame, but if an agent is so overcome with emotion that he cannot prevent himself from doing something morally good, there is no parallel tendency to give reduced praise. Rather, people tend to give him just as much praise as they would if he were able to maintain control.

Here, we suggest that this pattern, too, can be explained if one assumes that moral judgments are impacting ascriptions of the true self (for discussion, see Sripada, 2010). Specifically, in the case where the agent's emotions draw him to do something morally bad, we propose that these emotions are seen as lying outside his true self, and he is given less blame. However, in the case where the agent's emotions draw him to do something morally good, the emotions are seen as part of his true self, and he is given as much praise as if there were no conflict. So, for example, in the case where Jack is so overcome with sympathy that he gives a homeless man his jacket, we would predict that people will see this sympathy as part of his true self and will therefore assign him as much praise as if the action were deliberate. By contrast, in the case where Jack is so overcome with anger that he smashes a car window, people will not see this anger as part of his true self and will therefore assign him less blame than they would have if the action were deliberate. We tested this proposal in Experiment 4.

5.1. Method

Participants were 327 adults (M_{age} = 31.2, 55% female) who were recruited from mTurk and participated in exchange for $.25. Participants were assigned to one of 12 between-subjects conditions in a 2(moral valence: good vs. bad) × 2(action: deliberate vs. impulsive) × 3(vignette) design. The difference across vignettes was not hypothesized to be an important factor and served merely as a robustness check.

Participants read one of 12 vignettes (obtained from the authors) about an agent who, depending on condition, performed an action that was either good or bad and performed that action either deliberately or impulsively. The vignettes were very similar in structure and consisted of the following matched pairs: Man who (deliberately vs. impulsively) gives money away versus man who (deliberately vs. impulsively) steals; man who (deliberately vs. impulsively) gives away jacket versus man who (deliberately vs. impulsively) smashes a car window; male math teacher who (deliberately vs. impulsively) encourages female students versus a man who (deliberately vs. impulsively) fails to call on them.

Following each vignette, participants then rated their agreement with the following three statements: "How negatively or positively does this person deserve to be judged for their behavior? (1 = *extremely negatively*, 9 = *extremely positively*); How moral or immoral is this person's behavior? (1 = *extremely immoral*, 9 = *extremely moral*); What does this person deserve to receive for their behavior? (1 = *extreme blame*, 9 = *extreme praise*)." Following the procedure of Pizarro et al. (2003), these three measures were then averaged to form a composite scale for each vignette. These measures were then transformed so that both the praise and blame scores were reflected as positive values (i.e., ratings for the morally bad vignettes were reverse scored).

On a subsequent page, all participants rated their agreement (1 = *strongly disagree*, 9 = *strongly agree*) with two additional statements. One statement assessed beliefs about the true self: "To what extent do you think this action reflected (Bob's) true self—the person he really is deep down." In addition, because previous research (Pizarro et al., 2003) suggested a role of meta-desires, we also asked a question about whether the agent

wanted to have the impulse. For example, participants were asked, "To what extent do you think Bob wanted to have an impulse to give the elderly lady the extra $50?" The order in which each of these items appeared was counterbalanced across participants, and the agent's name was changed to match the particular vignette.

5.2. Results

A 2 (moral valence: good vs. bad) × 2 (action: deliberate vs. impulsive) × 3 (vignette) ANOVA revealed a marginal interaction between valence and action type on ratings of moral responsibility, $F(1,322) = 3.06$, $p = .08$. Replicating Pizarro et al. (2003), we observed that participants were significantly more likely to agree that the agent deserved moral blame when the bad action was deliberate ($M = 7.93$, $SD = 1.07$) than when the same action was impulsive ($M = 7.24$, $SD = 1.25$), $t(165) = 3.81$, $p < .001$. In contrast, there was no difference in ratings of moral praise between good behaviors that were performed deliberately ($M = 6.96$, $SD = 1.84$) versus those that were impulsive ($M = 6.66$, $SD = 1.85$), $p = .31$, ns. There was also a marginal three-way interaction with vignette type, $F(2,315) = 2.69$, $p = .07$. Inspection of the results across vignettes indicated that while the interaction effect was stronger for some vignettes than for others, the basic pattern of results was preserved across the different scenarios.

We then conducted a mediation analysis to determine whether beliefs about the true self explained the interaction of moral valence and action (deliberate vs. impulsive) on ratings of moral responsibility. We conducted a multiple-mediators analysis with the interaction term as the IV, ratings of moral responsibility as the DV, ratings of the true self and meta-desires as mediators, and the two main effects (valence and action) as covariates. This analysis indicated that beliefs about the true self significantly mediated the effect (95% CI = −.09 to −.01; see Fig. 5). In contrast, ratings of meta-desires did not (95% CI = −.01 to .02).

5.3. Discussion

This study replicated the results of Pizarro et al. (2003), where participants tend to assign equal praise to morally good actions that are performed either deliberately or impulsively, while for morally bad behavior participants assign more blame to deliberate actions than to impulsive actions. Following the logic of Experiments 1–3, we found that beliefs about the true self explain this effect. In the case where the agent's emotions draw him to do something morally bad, these emotions are seen as lying outside his true self and, in turn, he is given less blame. However, in the case where the agent's emotions draw him to do something morally good, the emotions are seen as part of his true self and so he is given as much praise as if there were no conflict.

Within this broader theoretical context, one might predict that impulsive actions themselves would cue different types of beliefs depending on the particular context. For example, Pizarro et al. (2003; as well as this study) find that impulsive actions receive *less* blame than deliberate actions, whereas Critcher, Inbar, and Pizarro (2013) find that

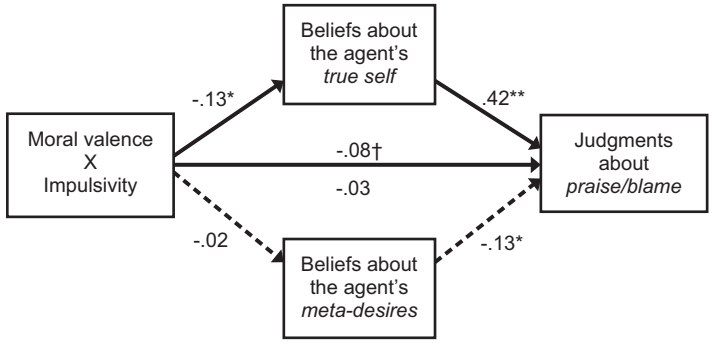

Fig. 5. Mediation results from Experiment 4. †$p = .10$; *$p < .01$; **$p < .001$.

agents who quickly decide to do harm are rated as *more* culpable than those that deliberate for longer and do harm. In other words, to the extent that impulsivity and quick decision making are related, people seem to have different judgments of moral responsibility for immoral impulsive actions depending on the particular context. On the present hypothesis, there is nothing about quick decision making in and of itself that makes an action be seen as more or less blameworthy. Rather, quick decision making can be framed in such a way that it is seen less as revealing an agent's true self (as in Pizarro et al., 2003) or in such a way that it is seen as more revealing of an agent's true self (as in Critcher et al., 2013). People's intuitions about blame then follow accordingly.

Finally, the results help to distinguish the role of judgments about the *true self* from the role of judgments about *meta-desires*. A long tradition of work in the philosophical literature, beginning with Frankfurt (1971), has closely identified these two notions, suggesting that an agent's true self is in some way determined by her meta-desires. Yet, in this study, we find these two notions coming apart, such that people's true self attributions systematically diverge from their meta-desire attributions (see also Newman et al., 2014, Study 3). More important, when the two notions are properly distinguished, we find that it is true self attributions—not meta-desire attributions—that explain the blame/praise asymmetry.

6. Experiment 5: Manipulating the true self

So far, our empirical approach across Experiments 1–4 has remained the same. For each experiment, we have used the identical stimuli from previous studies, added a measure related to the true self, and tested for mediation. While these experiments have produced a very consistent pattern of data—in each case, ratings of the true self appear to explain the previously identified asymmetries—the approach itself may be limited in that it does not directly test the causal effect of true self beliefs via experimental manipulation.

Thus, the goal of this study was to experimentally manipulate beliefs about the true self and measure the resulting effects on moral judgments. Specifically, participants were explicitly given information about an agent's true self (either good or evil). Then, they read the vignettes from the previous studies (i.e., on valuing, happiness, weakness of will, and praise/blame) and responded to the same dependent measures. Our prediction was that although participants may have a default to believe that the true self is good, if an agent is described as fundamentally evil, then this should affect subsequent morally relevant judgments. For example, participants should now be more likely to endorse the notion that an agent with an evil true self (vs. a good true self) is happy living a morally bad life.

Importantly, as in Experiments 1–4, we used the identical stimuli from previous research (adding only the relevant paragraph about the agent's true self, which appeared on a separate page). Note that this provides an extremely strict test of the hypothesis, as now our prediction is that the very same manipulation should produce the similar effects across stimuli coming from multiple topic areas, papers, laboratories, etc. Observing such an effect in this case would provide strong evidence for the notion that beliefs about the true self are in fact causally responsible for the observed asymmetries in moral judgments.

6.1. Method

Participants were 570 adults (M_{age} = 30.8, 33% female) who were recruited from mTurk and participated in exchange for $.50. Participants were assigned to one of eight between-subjects conditions in a 2(true self: good vs. evil) × 4 (judgment: valuing, happiness, weakness of will, praise/blame) design.

Participants initially read one of two stems about an agent named John. For half of the conditions, John was described as fundamentally evil, while in the other half of conditions he was described as fundamentally good. Participants read the following text (the alternate wording for each condition appears in brackets):

> Ever since John was born, it was clear that there was something distinctive about his personality. He sometimes did good [bad] things to other people, but deep down in his very essence, he was a fundamentally evil [good] person. At the very core of his being, he had no [a profound] compassion for other people and no concern at all [a genuine concern] about their well-being.
>
> When he was in his twenties, John went through a transitional phase. He was very confused about many things in his life and regularly abused drugs and alcohol. Most of the people he was spending time with were basically kind-hearted and friendly [pretty callous and cruel], but deep down within him, there was something that made him fundamentally different from all of them. On the next page you will read about an episode from that period in John's life.

> After that period, as John became an adult, he never felt the slightest amount [felt a great deal] of remorse about that particular time in his life. From that point on, he always pursued evil [good] in every way he could, and never expressed genuine concern [ill-will] toward another person. At his deepest level, he seemed to be fundamentally driven toward a life of bad [good].

As a manipulation check, participants then rated John's true self as 1 = *Fundamentally Evil* to 7 = *Fundamentally Good*. As expected, participants rated the agent's true self as significantly more evil when they read the "evil" stem (M = 5.82, SD = .99) than when they read the "good" stem (M = 1.87, SD = 1.20), $t(568) = 42.70$, $p < .001$.

Participants then read one of the 10 vignettes from Experiments 1–4, which consisted of one valuing vignette, four happiness vignettes, two weakness of will vignettes, and three praise/blame vignettes. The different vignettes always appeared between-subjects so that each participant only read one vignette. Consistent with our aim to limit degrees of freedom, for this study we used all of the scenarios from the previous studies, the identical text (with the exception that all of the agents were named John and the gendered statements were changed accordingly), and the same dependent measures.

Following the logic outlined in the introduction to this experiment, we tested vignettes in which the predicted effect stemming from an evil (vs. good) true self would result in greater endorsement of the mental state. Specifically, for valuing, the agent experienced an unwanted impulse toward discrimination and participants indicated whether the agent valued racial discrimination; for the happiness scenarios, the agent experienced happiness while living a "morally bad" life and participants indicated whether the agent was happy; and for the weakness of will scenarios the agent was resolved to keep doing bad things (murdering/stealing) but then lapsed, and participants indicated whether the agent showed weakness of will.

The praise/blame scenarios were somewhat different. In the initial study (Pizarro et al., 2003) and in the current Experiment 4, the only condition that differed from the others was the one in which the agent was so overcome with an impulse that he did something bad. In that case, participants indicated that the agent deserved less blame than if they deliberately performed the bad action. Therefore, we used the scenarios from the impulsive/bad conditions with the expectation that participants should determine that the agent deserved more blame when they were fundamentally evil versus fundamentally good.

6.2. Results

Where appropriate, the scores were reverse coded such that the predicted effect of the evil true self was equated with a larger scale value; that is, greater valuing, greater happiness, greater weakness of will, greater blame. We then collapsed the data within each type of vignette and normalized the scales for each of the respective judgments (i.e., valuing, happiness, etc.)

As our primary analysis, we conducted a 2(true self: good vs. evil) × 4(concept: valuing, happiness, weakness of will, praise/blame) ANOVA. This analysis indicated a

significant main effect of the true self, $F(1,562) = 8.96$, $p = .003$. As predicted, participants provided higher ratings (i.e., of valuing, happiness, weakness of will, and blame) when the agent was described as having an evil true self ($M = .15$, $SE = .07$) than when he was described as having a good true self ($M = -.14$, $SE = .07$).

In addition, there was no interaction between the manipulation of the true self (evil vs. good) and the type of concept, $F(3,561) = 1.15$, $p = .33$. Looking at each concept independently indicated that all were directionally consistent with the hypothesis (i.e., higher ratings when the agent's true self was evil vs. good). The effect sizes were as follows: Cohen's $d_{valuing} = .10$; Cohen's $d_{happiness} = .53$; Cohen's $d_{weaknessofwill} = .16$; Cohen's $d_{blame} = .35$.

6.3. Discussion

The results from this study are informative for several reasons. First, they provide direct evidence for the hypothesis that beliefs about the true self are in fact causally responsible for the observed asymmetries in moral judgments. It is important to note that in this study we again used the identical stimuli from previous research (adding only the relevant paragraph about the agent's true self, which appeared on a separate page). This provides an extremely strict test of the hypothesis as we found that the very same manipulation (literally the same paragraph) produced analogous effects across a wide variety of stimuli emanating from multiple topic areas, papers, laboratories, etc.

Second, these results address a possible alternative explanation of the original asymmetries. Looking at the original happiness asymmetry, for example, one might suggest that people simply do not want to say that a morally bad agent could be truly happy. (Perhaps the word "happy" is seen as serving a communicative function, conveying a certain level of approval of the agent's way of life.) Importantly, the present results show exactly the opposite of the pattern one would expect on that alternative hypothesis. Given that the agent performed a series of morally bad actions, participants actually rate him as *more* happy when he is described as fundamentally evil than when he is described as fundamentally good. This result suggests that the original asymmetry was indeed a reflection of people's understanding of the true self rather than just an attempt to avoid condoning certain behaviors by using particular words.

Third, the new Experiment 5 helps to identify boundary conditions by demonstrating that the previously identified effects may be attenuated to some degree when participants are explicitly told that the agent's true self is evil. Note, however, that in terms of absolute magnitude, the effects in that study were rather modest, suggesting that even in the case where an agent is described as being fundamentally evil, participants may still posit some good within. This is a phenomenon found in numerous works of fiction, for example, in Milton's *Paradise Lost*, where even Satan is portrayed as having something within him calling him to be reunited with God or—to give an example from a very different corner of our culture—in the conclusion of the *Star Wars* trilogy, where even Darth Vader is portrayed as having a hidden spark of compassion. The participants in this study

seem to be applying that same basic intuition to the description of the "evil" character in the vignettes. This may be an interesting issue to explore in future work.

7. General discussion

Across four studies, we found that beliefs in the true self appear to explain a number of asymmetries based on moral valence. Specifically, we examined the concepts of valuing (Experiment 1), happiness (Experiment 2), weakness of will (Experiment 3), and moral responsibility (Experiment 4). In each case, we found that beliefs about the true self explained participants' subsequent judgments, such that the moral asymmetries associated with each of these phenomena can all be explained by a more fundamental belief that "deep down" other agents are good. In a final study (Experiment 5), we directly manipulated beliefs about the true self and observed an effect on the application of each of the concepts.

These results are important and potentially surprising because on the surface, it would appear that concepts such as valuing, happiness, weakness of will, and praise/blame should have very little to do with the concept of a "true self." Nevertheless, the present studies indicate that each of these phenomena, which until now have been studied in isolation, can indeed be explained by a single underlying construct. We suggest that this research is theoretically valuable because it contributes to a growing interest to explain moral judgments in terms of more fundamental psychological processes. For example, Cushman and Young (2011) write, "Research in moral psychology faces the parallel challenge of distinguishing domain-specific moral computations from the effects of other domains on moral judgment. This is a challenge that echoes throughout the cognitive sciences as we discover how many 'higher' mental functions depend upon a core set of conceptual primitives, and thereby reflect their idiosyncratic structure" (p. 1070). While we find here that beliefs about the true self do appear to have a moral component (in assuming that the "true self" is fundamentally good), the present studies make significant headway by suggesting that many moral asymmetries may be embedded in richer conceptual structures used for reasoning about other agents' minds.

It is also worth noting that the present studies may be valuable from a methodological perspective. A potential worry in moral psychology (as well as more broadly) is that, given sufficient degrees of freedom, researchers can often "tailor" stimuli to find the desired effect (Simmons et al., 2011; Strickland & Suben, 2012). As a result, experiments may serve only to empirically reiterate intuition, rather than expose new phenomena. However, the present studies used stimulus materials that had already been developed within existing research (only adding relevant measures about the true self), thereby severely limiting degrees of freedom and providing a more conservative test of our hypotheses. Put differently, the original studies on valuing, happiness, weakness of will, and moral responsibility were not intended to examine the true self in any way; the fact that the true self does appear to "retrospectively" explain these effects provides quite strong support in favor of the primacy of the true self concept. Such an approach may

prove useful as researchers continue to integrate existing moral phenomena into larger psychological theories (cf. Cushman & Young, 2011).

7.1. A case study in moral judgment

We suggest that the present studies are best viewed as a case study in how beliefs about the true self can explain subsequent judgments (i.e., beliefs about valuing, happiness, weakness of will, and moral responsibility), and more broadly, how several different moral phenomena may be explained under a common theoretical umbrella. These four asymmetries were selected because they all seemed to be good candidates for testing the hypothesized role of the true self. However, this is not to say that this is an exhaustive list of all moral judgments/asymmetries that can be explained in this way—indeed, we think it is quite likely that there are others.

Recent work has found moral asymmetries in judgments about numerous matters in addition to the four discussed here. For example, there are asymmetries in people's judgments about *belief* (Leben & Wilckens, in press), *knowledge* (Beebe & Buckwalter, 2010), *agency* (Morewedge, 2009), *freedom* (Phillips & Knobe, 2009; Young & Phillips, 2011), and *causation* (Alicke, Rose, & Bloom, 2011). Future work could examine some of these other asymmetries to determine whether they too might be mediated by true self attributions.

A related series of articles by Sripada and colleagues (Sripada, 2010, 2012; Sripada & Konrath, 2011) have also argued that people's judgments about *intentional action* can be explained in terms of participants' judgments about the agent's self. Past research has identified an asymmetry in judgments of intentional action that in some ways resembles the asymmetries explored in the present article. Specifically, there is a tendency whereby participants are more likely to say that a side effect was brought about "intentionally" when that side effect is harmful than when it is helpful (Knobe, 2003). Sripada and colleagues show that the impact of condition (harm vs. help) on judgments about whether the agent acted intentionally is mediated by judgments about the agent's attitudes (Sripada & Konrath, 2011).

In some ways, this result looks similar to the patterns obtained in the present studies, but despite these apparent similarities, we think there are a number of reasons to suspect that the phenomena observed in the present studies are importantly different from the one observed in the studies on intentional action. Most notably, the present studies explicitly ask participants about how an agent feels "deep down"—that is, in the true self. By contrast, the studies on intentional action simply involved asking participants about an agent's general attitudes. For example, Sripada and Konrath (2011) asked participants to rate the agent on a scale from "very anti-environment" to "very pro-environment." The results showed that participants are more inclined to classify an action as intentional to the extent that this action is *concordant* with the agent's attitudes (e.g., that they are more inclined to regard the action of harming the environment as intentional to the extent that the agent is seen as anti-environmental; Sripada & Konrath, 2011). Importantly, however, asking about an agent's attitudes may be very different than asking about the agent's true

self (in fact, Experiments 1 and 2 in this article show precisely that). Therefore, it may be that while participants are likely to ascribe general attitudes that are concordant with an agent's immoral actions, they may nonetheless still be likely to posit a "deeper" self that is good. This seems like an interesting area for future research, which could look more closely at these two kinds of ascriptions and examine the differential effects they have on other aspects of cognition.

Regardless of what such an investigation would ultimately reveal, there is strong reason to suspect that the effects observed in the studies reported here are symptoms of a far broader phenomenon. It seems highly unlikely that people's way of thinking about the true self leads to exactly four asymmetries and all four of them have been uncovered in the present article. Rather, the natural assumption would be that there are a whole host of similar asymmetries to be discovered in further research.

7.2. Why is the true self "good"?

The principal aim of the present article was to offer an explanation for four moral asymmetries. We proposed that the four asymmetries could be explained in terms of a single underlying process, namely, people's tendency to see the true self as good. But this explanation immediately leads to a further question that could be addressed in future research. Perhaps the four asymmetries can be explained in terms of this single underlying process, but how are we to explain that process itself? Why *do* people have a tendency to think about the true self as good?

One possible explanation would be that this effect arises from something quite specific about the way people think about *human beings*. For example, the effect might reflect a "person-positivity bias" (Sears, 1983), whereby people tend to see human beings (but not objects of other types) in a positive light. Similarly, young children show a bias to believe that positive traits, such as good eyesight and intelligence, are retained through development while negative traits will spontaneously change in a positive direction—even biological negative traits such as poor eyesight and a missing finger (Lockhart, Chang, & Story, 2002). This might reflect an initial and general bias to think of persons in a positive way, so much so that when children are told about individuals who have negative traits, they believe that these will naturally disappear over time.

Alternatively, however, the effect might reflect something far more general about people's *psychological essentialism*, that is, their tendency to think of entities in terms of a deep underlying "essence" (e.g., Barsalou, 1985; Bloom, 2004, 2010; Gelman, 2003; Keil, 1989; Lynch, Coley, & Medin, 2000; Newman & Keil, 2008). Independent of anything about the way people think of human beings in particular, a growing body of evidence suggests that people tend to see essences as good. For example, just as people's values can shape their judgments about the "true self," it appears that 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). One might hypothesize, then, that the effect obtained for judgments about human beings in the present studies is just one symptom of a far more general tendency that can also be observed in people's judgments

about non-human objects. As one example, consider judgments about nations. If people believe that the United States has certain good qualities and certain bad ones, one might predict that they should show exactly the same response they showed when thinking about human beings in the present studies. That is, they should conclude that the essence of the United States—what the nation is "really about"—is revealed more by the good qualities than by the bad.

As the present studies are concerned exclusively with judgments about human beings, the data obtained here are not sufficient to decide between these two basic forms of explanation. This is an important topic for further research.

7.3. Conclusion

This article examined four apparently independent asymmetries and suggested that all four could be explained in terms of the same underlying psychological process, namely, attributions of a "true self." Future work in this area could seek greater breadth (by looking for yet further asymmetries driven by the same process) or greater depth (by trying to explain why people understand the true self in the way they do). Regardless of the precise form it takes, however, such work can proceed by examining these asymmetries not as four separate and unrelated effects, but as four symptoms of a single unified phenomenon: the tendency to assume that, deep down, others are morally good.

Acknowledgments

The authors thank Paul Bloom, Fiery Cushman, and Shaun Nichols for their helpful comments.

References

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, *108*, 670–696.

Aristotle. (350 BC/1985). *Nicomachean ethics*. [Trans. T Irwin]. Indianapolis, IN: Hackett Pub.

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.

Baumeister, R. F. (1991). *Meanings of life*. New York: Guilford.

Beebe, J. R. (2013). Weakness of will, reasonability, and compulsion. *Synthese*, *190*, 4077–4093.

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

Bloom, P. (2004). *Descartes' baby*. New York: Basic.

Bloom, P. (2010). *How pleasure works*. New York: Norton.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*, 308–315.

Cushman, F. A., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*, 1052–1075.

Diener, E., Scollon, C. N., & Lucas, R. E. (2003). The evolving concept of subject well-being: The multifaceted nature of happiness. *Advances in Cell Aging and Gerontology*, *15*, 187–219.

Doucet, M., & Turri, J. (2013). *Two non-psychological factors affecting ordinary attributions of weakness of will: Stereotypes and consequences*. Unpublished manuscript. University of Waterloo, Waterloo, ON.

Foucault, M. (1984). On the genealogy of ethics: An overview of work in progress. In P. Rabinow (Ed.), *The Foucault reader* (pp. 340–372). New York: Pantheon.

Frankfurt, H. G. (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, England: Oxford University Press.

Gide, A. (1902). *L'immoraliste*. Paris, France: Mercure de France.

Gonnerman, C. (2008). Reading conflicted minds: An empirical follow-up to Knobe and Roedder. *Philosophical Psychology*, *21*, 193–205.

Holton, R. (1999). Intention and weakness of will. *Journal of Philosophy*, *96*, 241–262.

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.

Kauppinen, A. (2006). Lovers of the good: Comments on Knobe and Roedder. In *First annual on-line philosophy conference*.

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

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

Kernis, M. H., & Goldman, B. M. (2006). A multi-component conceptualization of authenticity: Theory and research. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp. 283–357). New York: Academic Press.

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

Knobe, J., & Doris, J. (2010). Responsibility. In J. Doris and The Moral Psychology Research Group (Ed.), *The moral psychology handbook* (pp. 321–354). Oxford, England: Oxford University Press.

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

Knobe, J., & Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues*, *19*, 131–147.

Leben, D., & Wilckens, K. (in press). Pushing the intuitions behind moral internalism. *Philosophical Psychology*.

Lewis, D. K. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, *63*, 113–137.

Lockhart, K. L., 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 and Cognition*, *28*, 41–50.

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

Mele, A. (2010). Weakness of will and akrasia. *Philosophical Studies*, *150*, 391–404.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*, 535–545.

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

Newman, G. E., & Keil, F. C. (2008). Where is 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.

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

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review*, *3*, 320–322.

Phillips, J., Nyholm, S., & Liao, S. (in press). The good in happiness. *Oxford Studies in Experimental Philosophy*.

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*, 267–272.

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

Sartre, J.-P. (1958/2003). *Being and nothingness*. [Trans. H Barnes]. London: Routledge.

Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition*, *22*, 75–99.

Schimel, J., Arndt, J., Pyszczynski, T., & Greenberg, J. (2001). Being accepted for who we are: Evidence 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.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smith, M. (1994). *The moral problem*. Oxford, England: Blackwell.

Sousa, P., & Mauro, C. (2013). The evaluative nature of the folk concepts of weakness and strength of will. *Philosophical Psychology*, DOI:10.1080/09515089.2013.843057.

Sripada, C. S. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*, 159–176.

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, *48*, 232–238.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, *26*, 353–380.

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, *3*, 457–467.

Twain, M. (1885/1994). *The adventures of Huckleberry Finn*. New York: Dover.

Watson, G. (1975). Free agency. *Journal of Philosophy*, *72*, 205–220.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, *119*, 166–178.