This paper may be cited as Strohminger N., Newman, G., and Knobe, J. (in press). The True Self: A psychological concept distinct from the self. *Perspectives on Psychological Science*.

The True Self: A psychological concept distinct from the self

Nina Strohminger
Program in Cognitive Science and Department of Psychology
Yale University

Joshua Knobe
Program in Cognitive Science and Department of Philosophy
Yale University

George Newman
Yale School of Management
Yale University

A long tradition of psychological research has explored the distinction between characteristics that are part of the self and those that lie outside of it. Recently, a surge of research has begun examining a further distinction. Even among characteristics that are internal to the self, people pick out a subset as belonging to the *true self*. These factors are judged as making people who they really are, deep down. In this paper, we introduce the concept of the true self and identify features that distinguish people's understanding of the true self from their understanding of the self more generally. In particular, we consider recent findings that the true self is perceived as positive and moral, and that this tendency is actor-observer invariant and cross-culturally stable. We then explore possible explanations for these findings and discuss their implications for a variety of issues in psychology.

Keywords: self; true self; personal identity; moral self

If you open *The Handbook of Social Psychology* and flip to the index, you'll find that there are more than 60 separate entries listed under *self*—and that's not including the sub-entries (S. T. Fiske, Gilbert, & Lindzey, 2010). This is more than any other word in the index, aside from *social* (as in "social psychology"). The nature of the self is at the heart of psychology, driving many of its most urgent and profound questions.

Because this literature is so vast, the self can be conceptualized in many different ways. One popular way is to define the self in contrast with the environment that surrounds it. As Gilbert and Malone (1995) poetically put it,

> "the human skin [is seen] as a special boundary that separates one set of 'causal forces' from another. On the sunny side of the epidermis are the external or situational forces that press inward on the person, and on the meaty side are the internal or personal forces that exert pressure outward."
>
> (p. 21)

This epidermis-centric view captures something of what it means to be a self. It is useful for determining whether, for instance, we are touching Nina's leg or her bedsheets, and whether that object in the distance is our date or some shrubbery. Indeed, this tradition has uncovered a spate of classic and important effects in social psychology (e.g. Jones & Harris, 1967; Jones & Nisbett, 1971; Gilbert & Malone, 1995).

But as productive as this distinction has been, recent research suggests that people have another, quite different, way of thinking about the self. Even among characteristics that clearly lie 'inside the skin,' people further differentiate between features that are part of the "true self" and those that fall outside it. This distinction appears to have implications for a variety of psychological phenomena.

Our aim is to provide an overview of the emerging literature on the true self concept and show why it is critical to draw a distinction between the self and the true self, both for theory and practice.

The features of the true self

The self contains multitudes: it is a body and a mind, organs and thoughts, desires and intentions, whims

The authors would like to thank Rebecca Schlegel, Dan Bartels, Roy Baumeister, Sarah Molouki, Kerry McKenzie, Jesse Summers, and Jesse Chandler for helpful comments on an earlier draft of this manuscript.

and dispositions. Are all parts of the self equally self-like, or are certain parts especially essential?

A close examination of certain cases illustrates how the self concept, as a whole, differs from the true self. Consider the musical *Grease*, where Sandy sheds her goody-goody persona to become a leather-clad, pelvis-thrusting bad girl. Surely all this smokiness and gyration is Sandy. But just as surely, this is a performance designed to gain the approval of her peers, not the 'real' Sandy. Or witness Ebenezer Scrooge, who transforms from crotchety miser to joyful humanitarian in *A Christmas Carol*. Although both instantiations of Scrooge are Scrooge, readers are drawn to the conclusion that the latter Scrooge has discovered his true nature.

Less fancifully, we can reflect on real-life cases of inner conflict. When an addict experiences conflict about whether to stay sober or use drugs, both of these opposing desires occur within the self, but people tend to believe that one of these desires is more authentic than the other (Frankfurt, 1988; Arpaly & Schroeder, 1999). A similar conflict arises for conservative gay Christians, who are torn between homosexual impulses and the conviction that homosexuality is a sin (Newman, Knobe, & Bloom, 2014). Simply recognizing that both desires are part of the self is not fine-grained enough to capture the intuition that only a subset of desires belong to the true self.

The notion that some parts of the self are more authentic than others crops up frequently in psychological research, albeit under a slew of guises. Sometimes it is called the real self (Rogers, 1961; Turner, 1976; Masterson, 1988; Koole & Kuhl, 2003; Sloan, 2007), the ideal self (Chodorkoff, 1954; D. T. Kenny, 1956; Higgins, 1987), the authentic self (Johnson & Boyd, 1995; Cable, Gino, & Staats, 2013), the intrinsic self (Schimel, Arndt, Pyszczynski, & Greenberg, 2001; Arndt, Schimel, Greenberg, & Pyszczynski, 2002), the essential self (Strohminger & Nichols, 2014), or the deep self (Sripada, 2010).

However, a consensus is emerging over the term "true self", which we adopt here (Sheldon, Ryan, Rawsthorne, & Ilardi, 1997; Bargh, McKenna, & Fitzsimons, 2002; Johnson, Robinson, & Mitchell, 2004; Schlegel, Hicks, Arndt, & King, 2009; Newman, Lockhart, & Keil, 2010; Landau et al., 2011). While the inverse of true self is sometimes presented as "false self" (Harter, 2002; Johnson et al., 2004), the negation of true self as we intend it is a *superficial* or *peripheral* self—that is, aspects of the self that are inessential to who someone really is. The true self refers both to something we see in our own selves and in other people.

Historically, the true self concept has figured into psychological research in two rather different ways. Because the true self is a commonly-held belief amongst ordinary people, the bulk of scholarship has focused on describing how these beliefs work, and explicating their role in social behavior and cognition. But a subset of researchers make a bolder claim: the true self really does exist (Maslow, 1943; Rogers, 1961; Bem, 1973; Masterson, 1988; Koole & Kuhl, 2003; Kernis & Goldman, 2006). Carl Rogers, an influential proponent of this view, asserts that the true self lurks beneath the individual's "false front"; it is only "when [a person] fully experiences the feelings which at an organic level he *is*... that he is being a part of his real self" (Rogers, 1961, p. 111, emphasis original). (Not everyone agrees; Foucault's charmingly derisive term for Rogers and his ilk is "the Californian cult of the self"; 1983, p. 245.) Although in this paper we will treat the true self as a phenomenon of folk understanding, no doubt many readers will be wondering how these results bear on the ontological status of the true self. We will return to this question at the end of the paper.

Now that we have demarcated the pasture of the self landscape belonging to the true self, we can begin to chart its features. It may be helpful to consider the true self in contrast with the features that are commonly granted to the self (Table 1).

The true self is moral

A well-established finding within cognitive psychology is that changing the central features of a concept alters its identity more than changing its peripheral features (Sloman, Love, & Ahn, 1998). Thus, one way of discovering what the true self consists of is to test what kinds of changes to the self alter a person's identity the most. Broadly speaking, one's mind contributes to personal identity more than one's body (Blok, Newman, Behr, & Rips, 2001; Nichols & Bruno, 2010). But when psychological characteristics are pitted directly against one another (e.g. perception, memories, preferences, personality), people report the greatest identity discontinuity when *moral* capacities have been altered or removed (Strohminger & Nichols, 2014; Prinz & Nichols, in press). This pattern is quite robust. It shows up regardless of the source of change (the aging process, medical interventions, supernatural events), and regardless of the type of moral feature (disposition, behavior, or belief) (Heiphetz, Strohminger, & Young, in press; Molouki & Bartels, in press). Moral traits are considered to be the most deeply rooted, causally central aspect of a person's identity (Chen, Urminsky, & Bartels, in press)., which may help explain why people are unwilling to take psychopharmaceuticals that address moral deficits (Riis, Simmons, & Goodwin, 2008).

The privileging of moral traits within the self has

Table 1
A comparison of key differences between the true self and the self more generally.

THE SELF	THE TRUE SELF
Encompasses entire range of personal features	Emphasizes moral features
Valence-independent; can be positive or negative	Valence-dependent; positive by default
Perspective (first- or third-person) dependent	Perspective-independent
Cross-culturally variable	Cross-culturally stable

also been observed in real cases of neurological change. Patients with frontotemporal dementia—a disease that primarily disrupts moral capacities—are seen by loved ones as having changed more "deep down" than patients with diseases that primarily impact memory (Alzheimer's) or motor function (amyotrophic lateral sclerosis). Even in neurodegenerative diseases that primarily impact other parts of the mind, the extent to which patients are perceived as essentially the same person is almost entirely predicted by whether their moral capacities have been preserved (Strohminger & Nichols, 2015).

The dominance of moral traits in personal identity appears to extend to impression formation as well. While it has long been recognized that warmth matters more than competence in forming impressions of others (Anderson, 1968; Wojciszke, Bazinska, & Jaworski, 1998), more recently it has been discovered that morality plays a larger role than either (Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Goodwin, Piazza, & Rozin, 2014; Goodwin, 2015). This commonality suggests is a deeper level at which people perceive others: not just as a loose conglomeration of personality traits, but also, and perhaps especially, in terms of their underlying moral selves.

This is not to say that the true self only comprises moral features. People also attribute emotions, desires, and hidden mental states to the true self (Andersen & Ross, 1984; Bench, Schlegel, Davis, & Vess, 2015). Nonetheless, current research suggests that moral features play a uniquely constitutive role.

The true self is good

The true self is not just perceived as moral, but as good. Moral improvement leads to less perceived personal identity discontinuity than moral deterioration (Tobia, 2016; Molouki & Bartels, in press). Generally speaking, positive personal changes are seen as discoveries. That is, they are not seen as a form of change at all, but as revealing what was always hidden deep inside (Bench et al., 2015). This may explain why the feeling of knowing who someone really is deep down is strongest when subjects are given both moral and positive information about a target (Christy, Kim, Schlegel, Vess, & Hicks, 2016). It may also explain why mental illness is often portrayed in clinical psychology as 'covering up' the real self (Masterson, 1988).

This view of the true self as an underlying, and potentially invisible, aspect of the self is bolstered by research on psychological essentialism. Positive, desirable personality traits are more essentialized than negative, undesirable traits, and essentialized traits are in turn seen as more central to defining to who someone is (Haslam, Bastian, & Bissett, 2004). Adults and children share the intuition that a person's traits will tend to improve over time, indicating an implicit belief that, regardless of its current surface features, the true self's positive nature will eventually shine through (Lockhart, Chang, & Story, 2002; Molouki & Bartels, in press). Importantly, converging evidence now suggests that there is an independent contribution of positive valence: the more positive a trait is—moral or not—the more likely it is seen as part of the true self (e.g. Bench et al., 2015; Molouki & Bartels, in press; Tobia, 2016; Christy, Kim, et al., 2016).

The importance of positive valence to the true self becomes even clearer in studies where subjects must choose whether features belong to the true self or the surface self. When asked which part of the self is responsible for a person becoming bad (e.g. a deadbeat dad), subjects attribute this change to the surface self, but becoming a better person (e.g. a loving father) is attributed to the true self (compare this finding to Scrooge's transformation; Newman, Knobe, & Bloom, 2014). This effect is contingent on the values of the person rendering the judgment: liberals think homosexual urges are part of the true self, but conservatives think it is not (Newman, Knobe, & Bloom, 2014). And though there is a tendency to consider feelings to be more self-like than cooler cognitive states (Haslam et al., 2004; Johnson et al., 2004), whether feelings or beliefs are considered part of the true self depends on whether they

are good (Newman, Knobe, & Bloom, 2014).

Even misanthropes and pessimists believe that the true self is good (De Freitas et al., 2016). To date, the only manipulation that has been able to eliminate this tendency is one in which participants are explicitly told that an agent has a morally bad true self (Newman, De Freitas, & Knobe, 2014). Even then, participants' tendency to see the agent's true self as morally bad is not as strong as the tendency to see the true self as morally good when participants are given no instruction at all. Though we are perfectly willing to conceive of other people as bad, we are unwilling to see them as bad deep down.

Taken together, these findings demonstrate an important respect in which people's understanding of the true self is different from their understanding of the self. People are happy to regard morally bad motives as internal to a person (e.g. S. T. Fiske, 1980; Wojciszke, Brycz, & Borkenau, 1993), but they are nonetheless reluctant to regard such motives as elements of a person's true self. Existing studies point to a strong tendency for people to see such motives as external to the true self and to conclude that the true self is calling the person to do what is morally right.

The true self is perspective-independent

One of the most ubiquitous findings in social psychology is that the assessments people make for themselves differ from the assessments they render unto others. When judging others, negative information is more salient and powerful than positive information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). Our impressions of others are not simply the sum of all good and bad traits, as negative traits are weighted more heavily (S. T. Fiske, 1980; Wojciszke et al., 1993). Less evidence is required to designate someone a sinner than a saint, and relatively small amounts of sinning can relegate one to the sinner category (Reeder & Coovert, 1986; Klein & O'Brien, in press). People are, indeed, all too eager to attribute the unsavory behaviors of the people around them to internal causes, even when these behaviors can be clearly traced to random, situational causes (Jones & Harris, 1967; Gilbert & Malone, 1995).

When people judge themselves, this pattern reverses. People routinely overestimate their own knowledge and abilities (Gilovich, 1991; West & Stanovich, 1997; Kruger & Dunning, 1999; Rozenblit & Keil, 2002) and downplay the role of situational factors when accounting for their own accomplishments (Jones & Nisbett, 1971; Malle, 2006). This self-serving bias leads people to deem themselves superior to others in virtually every testable way— more virtuous, more skilled, more rational, more unique (Alicke, 1985; Brown, 1986; Falk, 1989; Kenworthy & Miller, 2002; Klein & Epley, 2016). Naturally, people also believe themselves to be less prone to cognitive biases than the average schmo (Pronin, Lin, & Ross, 2002; Pronin, Gilovich, & Ross, 2004). Note, however, that the bulk of research comparing own–other attributions has been concerned with the "self" at a general level.

When subjects are instead asked about the true self, no such actor-observer asymmetry emerges. People regard their own true selves as good (Bench et al., 2015; Molouki & Bartels, in press), but they also consider the true selves of others to be fundamentally good (Bench et al., 2015; Newman, Knobe, & Bloom, 2014) and moral (Heiphetz et al., in press). Unlike more global assessments of the self, the true self is painted in flattering colors whether one is looking inward, at one's own self, or outward, to others. There may well be exceptions at the individual level—perhaps psychopaths seem deeply flawed only to observers, and vice versa for the clinically depressed. There also seem to be some perspective-based differences in the kinds of content attributed to the true self, with people believing that experiences best reflect their own true selves, whereas general dispositions best reflect others' (Johnson & Boyd, 1995). Nonetheless, perspective-independence in valence is the pattern that obtains at the population level.

It is worth emphasizing just how striking this discrepancy is. One of the most ubiquitous effects in the self literature—actor-observer valence asymmetry—fails to obtain for true self attribution. This regularity gives us a hint about the process underlying true self attribution, which we shall return to shortly.

The true self is cross-culturally stable

A large body of work has been devoted to how conceptions of the self differ cross-culturally. In very broad strokes, Westerners understand the self in terms of the individual (*independence*), whereas Easterners understand the self more in terms of social relationships (*interdependence*; Markus & Kitayama, 1991).

Because of such differences in the way the self is conceptualized across Eastern and Western cultures, it seems natural to expect differences in the true self. However, the true self concept is remarkably cross-culturally robust. Like Westerners, people from a range of interdependent cultures (Russia, Singapore, and Colombia) hold the belief that the true self is normatively good (De Freitas et al., 2016). Similarly, Americans and Japanese believe that negative traits (such as being unkind or having bad vision) will tend 'correct themselves' over time, belying an im-

plicit belief that the positive trait represents the person's deeper nature (Lockhart, Nakashima, Inagaki, & Keil, 2008). The moral dimension of the true self is also preserved cross-culturally: Hindu Indians and Buddhist Tibetans consider moral features more central to personal essence than any other psychological trait (Nichols, Strohminger, Rai, & Garfield, 2016). This finding is particularly striking given that Buddhist Tibetans explicitly deny the existence of the self (Garfield, Nichols, Rai, & Strohminger, 2015).

Based on the available evidence, many of the key features of the true self—that it is moral, good, and perspective-independent—are culturally invariant. Of course, the present account also predicts a specific kind of cultural difference. In particular, when different cultures have different views about what is morally good, these different cultures should show correspondingly different patterns of true self attributions. However, this predicted cross-cultural difference would reveal a similarity at a more abstract level. Different cultures might have different views about which actions are morally good, but they could be similar in believing that the true self is calling us to morally good actions.

Directions for future research

Consequences of the true self concept

Thus far, we have been suggesting that it can be helpful to distinguish between the way people understand the self in general and way people understand the true self in particular. We turn now to some possible downstream effects of this distinction. Some of these effects have already been examined in existing research, while others still remain to be explored.

Within research on well-being, studies show that judgments about the true self have a special connection to people's sense of meaning in life. Numerous studies have shown that people generally hold positive attitudes about the self (e.g. Gilovich, 1991; Kruger & Dunning, 1999); however, the effect on people's sense of meaning in life is specific to attitudes about the true self (Schlegel et al., 2009; Schlegel, Hicks, King, & Arndt, 2011). Suppose that a person has a desire to make a lot of money and also a desire to create a beautiful work of art. This person may see both desires as aspects of her self, but to the extent that she sees only the latter as falling within her true self, the satisfaction of this latter desire will contribute to her sense of meaning in life in a way that the satisfaction of the former will not. Research has also found a distinctive impact of true self beliefs on a number of other outcomes, including satisfaction with life decisions (Schlegel, Hicks, Davis, Hirsch, & Smith, 2013; Kim, Christy, Hicks, & Schlegel, 2016), feelings of defensiveness (Schimel et al., 2001), and motivation and well-being (Ryan & Deci, 2000).

Within clinical research, the distinction between self and true self may shed light on the reluctance people show to pursue courses of action that would lead to changes in the self. Work by Riis and colleagues (2008) found that people were reluctant to use psychopharmaceuticals to enhance aspects of the self that were seen as lying at the core of their identities (e.g. empathy, kindness, mood), but they were perfectly willing to adopt the same approach to enhancing aspects of the self that were seen as peripheral to their identities (e.g. concentration, memory). Future studies could ask whether this same effect arises for other treatment modalities. For example, people with clinical depression and anxiety disorders are often reluctant to seek treatment, even when the treatment in question is talk therapy (Kohn, Saxena, Levav, & Saraceno, 2004; Ilse van Beljouw et al., 2010; Mojtabai et al., 2011; Nyholm & O'Neill, in press). This reluctance may arise in part because people see their emotional state as an expression of their true selves (Kramer, 1993).

Within moral cognition, existing research finds an asymmetry in people's judgments about blame versus praise for actions that were driven by overwhelming, irresistible emotion. Agents receive decreased blame for morally bad actions when overcome by emotion, but they do not receive decreased praise for morally good actions when overcome by emotion (Pizarro, Uhlmann, & Salovey, 2003). Incorporating the true self concept can provide a straightforward explanation of this effect. Although emotion is seen as part of the self in both cases, people should be inclined to see it as falling outside the true self in the morally bad case (leading to decreased blame) but as falling inside the true self in the morally good case (leading to no decrease in praise). Recent studies find support for this explanation using both mediation and manipulation (Newman, De Freitas, & Knobe, 2014). Moreover, subsequent work suggests that intuitions about the true self might also explain a variety of other patterns in blame judgments, including the tendency to decrease blame in cases where the agent had a bad upbringing (Faraci & Shoemaker, 2016) or mental illness (Daigle & Demaree-Cotton, 2016), or in cases where a large amount of time has elapsed (Mott, 2016).

Within the study of relationships, extant work shows that people employ different norms depending on the type of relationship they are in (A. P. Fiske, 1992; Clark & Aragon, 2013). For instance, paying for services is appropriate in the context of a market pricing relationship, but it would be in poor taste for a friendship. Perhaps one such difference is that people are inclined to

think about the true self in certain kinds of relationships (e.g. romantic) but are less likely to consider the true self in others (e.g. exchange-based relationships). If this hypothesis turns out to be correct, it could help explain the well-documented bias wherein people to evaluate their romantic partners less accurately, and more favorably, than strangers do (D. A. Kenny & Acitelli, 2001; Gagné & Lydon, 2004; Geher et al., 2005). Under this view, people in romantic relationships do not simply arrive at different answers; they are asking fundamentally different questions. While people in most relationships aim to understand the entirety of the person's traits, it might be that people in romantic relationships focus especially on describing their partner's true self.

Finally, the distinction may help in understanding how to alleviate intergroup conflict. Intergroup conflict is often fueled by attributions involving the self (Pettigrew, 1979; Hewstone, 1990). Conflicts between groups arise and are intensified by differential attitudes that outgroup members take toward political issues, controversial moral questions, musical and fashion choices, even arbitrary group assignment (Allen & Wilder, 1975; Tajfel & Turner, 1979/2001; Locksley, Ortiz, & Hepburn, 1980; Lonsdale & North, 2009). Emerging evidence suggests that such conflicts could be alleviated to the degree that people instead focus on outgroup members' true selves (De Freitas & Cikara, 2016). While people might continue to acknowledge that the outgroup members differ from themselves on a variety of dimensions, reminding them that outgroup members share a common features at a deeper level reduces intergroup bias.

Is the true self always perceived as morally good?

Given how much variation there is in personal traits—both among judgers and the judged—it may come as a surprise that there is such uniformity in assessments of the true self. People with predominantly bad surface traits are generally considered to be good deep down, just like people who have predominantly good surface traits (Newman, Knobe, & Bloom, 2014). Likewise, misanthropy and pessimism scores add no predictive value to the tendency to attribute goodness to others' true selves (De Freitas et al., 2016). But, like all good psychological generalizations, this one comes with a few provisos.

First, there appear to be individual differences in the attributes assigned to the true self. People with above average psychopathy scores do not consider morality to be the most important trait when judging the identity continuity of other people, and those very low in psychopathy place an especially strong emphasis on morality (Strohminger & Nichols, 2016). We might expect other individual differences to yield variations in the traits attributed to the true self, such as theory of mind capabilities (viz. Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) or autobiographical memory function (viz. Palombo, Williams, Abdi, & Levine, 2013).

The underlying rule that would explain such differences is observers ascribe to the true self traits that they themselves value. Psychopaths don't value being morally good, therefore changes to the moral features do not register as fundamentally disruptive of identity. Misanthropists still value being morally good, they just have grown weary and skeptical. A cynic is a disappointed idealist, but a psychopath is no idealist at all.

Second, although people have a strong tendency to attribute morally good desires to the true self, people will also attribute morally bad desires to the true self under the right circumstances. For instance, a man who morally endorses same-sex couples but is nonetheless repulsed by them is seen as fundamentally conflicted. For many subjects, both his automatic reaction and the principled belief are deemed part of his true self (Newman, Knobe, & Bloom, 2014, Study 3). The existence of such cases indicates that people do not always see the true self as a unitary phenomenon. Sometimes they posit competing impulses within it.

Why are true self attributions this way?

Existing research suggests a striking convergence in peoples true self attributions. Regardless of culture, perspective or personality, people tend to see the true self as morally good. Why is this the case? We see two possible types of explanation, corresponding to the traditional distinction between "motivational" and "cognitive" approaches (Kunda, 1990).

The first possibility is that the effects we have reviewed thus far can be explained in terms of motivated cognition. At the core of the motivated cognition approach is the idea that people hold certain views because they experience an external need or desire to hold those views. Numerous studies in other domains have found motivational effects of this type (Jost, Banaji, & Nosek, 2004; Kunda, 1990), and there is strong evidence that motivated cognition plays an important role in person perception (Kuzmanovic, Jefferson, Bente, & Vogeley, 2013; Lemay Jr., 2014). Quite plausibly, this same basic process is at work in true self attributions. Perhaps people believe that all human beings are morally good deep down because there is an external benefit to doing so, such as enhancing well-being or interpersonal trust.

An alternative possibility is that true self attributions arise as a result of more domain-general cognitive processes. In support of this view, recent work finds that a wide range of non-human entities are also seen as essentially good. Nations, rock bands, universities, scientific papers, and conferences are judged to retain more of their identity when their features improve than when they deteriorate (De Freitas, Tobia, Newman, & Knobe, in press). It seems unlikely that people are motivated to believe that, deep down, all conferences are good. Rather, this tendency suggests that value judgments truly are playing a role in peoples reasoning.

One way to spell out this second possibility is that judgments about the true self are a product of *psychological essentialism* (Barsalou, 1985; Keil, 1989; Lynch, Coley, & Medin, 2000; Gelman, 2003; Bloom, 2004, 2010; Newman & Keil, 2008). Consistent with this view, recent studies show that beliefs about the true self are characterized by telltale features of essentialist reasoning, such as immutability, informativeness, and inherence (Christy, Schlegel, & Cimpian, 2016). The effects we have been reviewing here might then be explained in terms of more general facts about how psychological essentialism works. For example, independent of how people think of human beings, a growing body of evidence suggests that people tend to see essences as good (Knobe, Prasada, & Newman, 2013; De Freitas et al., in press). The propensity to see others as having a good true self might then be explained in terms of this quite general fact about the way people attribute essences. One advantage of this view is it would explain the absence of a perspective effect. The shift between actor and observer perspectives leads to a substantial difference in motivation, but it might have no effect at all on the basic processes involved in essentialist reasoning.

One remaining puzzle is how to explain the finding that people specifically see the true self as *morally* good rather than good in some other way. The exact reason for this is open at the moment. Previous work shows that an object's identity is related to its purpose (Gelman & Bloom, 2000; Rose & Schaffer, in press). Perhaps the same is true of people, with morality being seen as the human *telos*. Another, non-mutually exclusive, possibility is that moral traits are especially important because they are essential for the maintenance of social bonds (Strohminger & Nichols, 2015; Heiphetz et al., in press). Because humans are hyper-social creatures, they may be keenly tuned to tracking the moral features of the people around them.

While research shows that the true self is *principally* moral, it also shows that many valued traits are ascribed to the true self (Bench et al., 2015). In American culture, the idea that all of us have within us a trove of hidden talents and abilities waiting to be exposed looms large, and its popularity may be made possible by the notion of the true self.

Coda: Does the true self exist?

In this paper, we have outlined the principal features of a folk concept, and discussed the role it plays in various aspects of human thought and behavior. Readers may now be curious about a deeper question. Is the true self also a *scientific* concept, one that can be used to describe how the mind actually works? Is there, in other words, a true self?

The evidence reviewed here points to two properties relevant to this question. One: the true self depends on the values of the observer. If someone thinks homosexual urges are wrong, she will say the desire to resist such urges represents the true self (Newman, Knobe, & Bloom, 2014). And if she scores high in psychopathy, she will assign less weight to moral features in her conceptualization of personal identity (Strohminger & Nichols, 2016). What counts as part of the true self is subjective, and strongly tied to what each individual person herself most prizes.

Two: The true self is, shall we say, evidence-insensitive. Resplendent as the true self is, it is also a bashful thing. Yet people have little trouble imbuing it with a host of hidden properties. Indeed, claims made on its behalf may completely contradict all available data, as when the hopelessly miserable and knavish are nonetheless deemed good 'deep down'. The true self is posited rather than observed. It is a hopeful phantasm.

These two features—radical subjectivity and unverifiability—prevent the true self from being scientific concept. The notion that there are especially authentic parts of the self, and that these parts can remain cloaked from view indefinitely, borders on the superstitious. This is not to say that lay belief in a true self is dysfunctional. Perhaps it is a useful fiction, akin to certain phenomena in religious cognition and decision-making (Boyer, 2001; Gigerenzer & Todd, 1999). But, in our view, it is a fiction nonetheless.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630.

Allen, V. L., & Wilder, D. A. (1975). Categorization, belief similarity, and intergroup discrimination. *Journal of Personality and Social Psychology*, *32*(6), 971–977.

Andersen, S. M., & Ross, L. (1984). Self-knowledge and social inference: The impact of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology*, *46*(2), 280–293.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*(3), 272–279.

Arndt, J., Schimel, J., Greenberg, J., & Pyszczynski, T. (2002). The intrinsic self and defensiveness: Evidence that activating the intrinsic self reduces self-handicapping and conformity. *Personality and Social Psychology Bulletin*, *28*(5), 671–683.

Arpaly, N., & Schroeder, T. (1999). Praise, blame and the whole self. *Philosophical Studies*, *93*(2), 161–188.

Bargh, J. A., McKenna, K. Y., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the internet. *Journal of Social Issues*, *58*(1), 33–48.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251.

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*(4), 629–654.

Baumeister, R., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Bem, D. J. (1973). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. Vol. 6, pp. 1–62). New York, NY: Academic Press.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, *33*(3), 169–185.

Blok, S. V., Newman, G., Behr, J., & Rips, L. J. (2001). Inferences about personal identity. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 80—85). Mahwah, NJ: Erlbaum.

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

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

Boyer, P. (2001). *Religion explained*. New York, NY: Basic Books.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition*, *4*(4), 353–376.

Cable, D. M., Gino, F., & Staats, B. R. (2013). Breaking them in or eliciting their best? Reframing socialization around newcomers' authentic self-expression. *Administrative Science Quarterly*, *58*(1), 1–36.

Chen, S. Y., Urminsky, O., & Bartels, D. M. (in press). Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science*.

Chodorkoff, B. (1954). Adjustment and the discrepancy between the perceived and ideal self. *Journal of Clinical Psychology*, *10*(3), 266–268.

Christy, A., Kim, J., Schlegel, R., Vess, M., & Hicks, J. (2016). *Do I know you? Moral information predicts perceived knowledge of others' true selves.* (Unpublished manuscript.)

Christy, A., Schlegel, R., & Cimpian, A. (2016). *Why do people believe in true selves? The role of psychological essentialism.* (Unpublished manuscript.)

Clark, M. S., & Aragon, O. (2013). Communal (and other) relationships: History, theory development, recent findings, and future directions. In J. A. Simpson & L. Campbell (Eds.), *The oxford handbook of close relationships* (pp. 255–280). New York, NY: Oxford University Press.

Daigle, J., & Demaree-Cotton, J. (2016). *One reason why normative competence theorists should avoid the case strategy.* (Unpublished manuscript.)

De Freitas, J., & Cikara, M. (2016). *Deep down my enemy is good: Thinking about the true self reduces intergroup bias.* (Unpublished manuscript.)

De Freitas, J., Sarkissian, H., Grossman, I., De Brigard, F., Luco, A., Newman, G., & Knobe, J. (2016). *Is there universal belief in a good true self?* (Unpublished manuscript.)

De Freitas, J., Tobia, K., Newman, G., & Knobe, J. (in press). The good ship Theseus: The effect of valence on object identity judgments. *Cognitive Science*.

Falk, R. (1989). Judgment of coincidences: Mine versus yours. *The American Journal of Psychology*, *102*(4), 477–493.

Faraci, D., & Shoemaker, D. (2016). *Good selves, true selves.* (Unpublished manuscript.)

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689–723.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906.

Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). *Handbook of social psychology* (5th ed., Vol. 2). Hoboken, NJ: John Wiley & Sons.

Foucault, M. (1983). On the genealogy of ethics: An overview of work in progress. In H. Dreyfus & P. Rabinow (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics* (2nd ed.). Chicago, IL: University of Chicago Press.

Frankfurt, H. G. (1988). *Freedom of the will and the concept of a person*. New York, NY: Springer.

Gagné, F. M., & Lydon, J. E. (2004). Bias and accuracy in close relationships: An integrative review. *Personality and Social Psychology Review*, *8*(4), 322–338.

Garfield, J. L., Nichols, S., Rai, A. K., & Strohminger, N. (2015). Ego, egoism and the impact of religion on ethical experience: What a paradoxical consequence of buddhist culture tells us about moral psychology. *The Journal of Ethics*, *19*(3-4), 293–304.

Geher, G., Bloodworth, R., Mason, J., Stoaks, C., Downey, H. J., Renstrom, K. L., & Romero, J. F. (2005). Motivational underpinnings of romantic partner perceptions: Psycho-

logical and physiological evidence. *Journal of Social and Personal Relationships*, *22*(2), 255–281.

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

Gelman, S., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, *76*(2), 91–103.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford, England: Oxford University Press.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38.

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

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168.

Harter, S. (2002). Authenticity. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 382–394). New York, NY: Oxford University Press.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, *30*(12), 1661.

Heiphetz, L., Strohminger, N., & Young, L. (in press). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science*.

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, *20*(4), 311–335.

Higgins, E. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*(3), 319–340.

Ilse van Beljouw, M. S. B., Verhaak, P., Prins, M., Cuijpers, P., Penninx, B., Bensing, J., et al. (2010). Reasons and determinants for not receiving treatment for common mental disorders. *Psychiatric Services*, *61*(3), 250–257.

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*(4), 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*(5), 615.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Jones, E. E., & Nisbett, R. E. (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*(6), 881–919.

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

Kenny, D. A., & Acitelli, L. K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology*, *80*(3), 439–448.

Kenny, D. T. (1956). The influence of social desirability on discrepancy measures between real self and ideal self. *Journal of Consulting Psychology*, *20*(4), 315–318.

Kenworthy, J. B., & Miller, N. (2002). Attributional biases about the origins of attitudes: Externality, emotionality and rationality. *Journal of Personality and Social Psychology*, *82*(5), 693–707.

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

Kim, J., Christy, A., Hicks, J. A., & Schlegel, R. J. (2016). *Trust thyself: True-Self-as-Guide lay theories enhance decision satisfaction.* (Unpublished manuscript.)

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology*, *110*(5), 660–674.

Klein, N., & O'Brien, E. (in press). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition*.

Knobe, J., Prasada, S., & Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257.

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World Health Organization*, *82*(11), 858–866.

Koole, S. L., & Kuhl, J. (2003). In search of the real self: A functional perspective on optimal self-esteem and authenticity. *Psychological Inquiry*, *14*(1), 43–48.

Kramer, P. (1993). *Listening to Prozac*. New York, NY: Viking.

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*(6), 1121–1134.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kuzmanovic, B., Jefferson, A., Bente, G., & Vogeley, K. (2013). Affective and motivational influences in person perception. *Frontiers in Human Neuroscience*, *7*, 266.

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*(1), 79–87.

Lemay Jr., E. P. (2014). Accuracy and bias in self-perceptions of responsive behavior: Implications for security in romantic relationships. *Journal of Personality and Social Psychology*, *107*(4), 638–656.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism? *Child Development*, *73*(5), 1408–1430.

Lockhart, K. L., Nakashima, N., Inagaki, K., & Keil, F. C. (2008). From ugly duckling to swan?: Japanese and amer-

ican beliefs about the stability and origins of traits. *Cognitive Development*, *23*(1), 155–179.

Locksley, A., Ortiz, V., & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology*, *39*(5), 773–783.

Lonsdale, A. J., & North, A. C. (2009). Musical taste and ingroup favouritism. *Group Processes & Intergroup Relations*, *12*(3), 319–327.

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*(1), 41–50.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*(6), 895–919.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

Masterson, J. F. (1988). *The real self: A developmental, self and object relations approach*. New York, NY: Simon & Schuster.

Mojtabai, R., Olfson, M., Sampson, N. A., Jin, R., Druss, B., Wang, P. S., … Kessler, R. C. (2011). Barriers to mental health treatment: Results from the National Comorbidity Survey Replication. *Psychological Medicine*, *41*(8), 1751–1761.

Molouki, S., & Bartels, D. M. (in press). Personal change and the continuity of identity. *Cognitive Psychology*.

Mott, C. (2016). *Statutes of limitations and personal identity.* (Unpublished manuscript.)

Newman, G., De Freitas, J., & Knobe, J. (2014). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125.

Newman, G., & Keil, F. C. (2008). Where is the essence? Developmental shifts in children's beliefs about internal features. *Child development*, *79*(5), 1344–1356.

Newman, G., Knobe, J., & Bloom, P. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203–216.

Newman, G., Lockhart, K. L., & Keil, F. C. (2010). Ş̈end-of-lifeŤ biases in moral evaluations of others. *Cognition*, *115*(2), 343–349.

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

Nichols, S., Strohminger, N., Rai, A. K., & Garfield, J. L. (2016). *Death and the self.* (Unpublished manuscript.)

Nyholm, S., & O'Neill, E. (in press). Deep brain stimulation, continuity over time, and the true self. *Cambridge Quarterly of Healthcare Ethics*.

Palombo, D. J., Williams, L. J., Abdi, H., & Levine, B. (2013). The survey of autobiographical memory (sam): A novel measure of trait mnemonics in everyday life. *Cortex*, *49*(6), 1526–1540.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*(4), 461–476.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise the role of perceived metadesires. *Psychological Science*, *14*(3), 267–272.

Prinz, J., & Nichols, S. (in press). Diachronic identity and the moral self. In J. Kiverstein (Ed.), *Handbook of the social mind.* London: Routledge.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.

Reeder, G. D., & Coovert, M. D. (1986). Revising an impression of morality. *Social Cognition*, *4*(1), 1–17.

Riis, J., Simmons, J. P., & Goodwin, G. P. (2008). Preferences for enhancement pharmaceuticals: The reluctance to enhance fundamental traits. *Journal of Consumer Research*, *35*(3), 495–508.

Rogers, C. R. (1961). *On becoming a person: A therapist's view of psychology*. Boston, MA: Houghton Mifflin.

Rose, D., & Schaffer, J. (in press). Folk mereology is teleological. *Noûs*. doi: 10.1111/nous.12123

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

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*(1), 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*(2), 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*(3), 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*(6), 745–756.

Sheldon, K. M., Ryan, R. M., Rawsthorne, L. J., & Ilardi, B. (1997). Trait self and true self: Cross-role variation in the big-five personality traits and its relations with psychological authenticity and subjective well-being. *Journal of Personality and Social Psychology*, *73*(6), 1380–1393.

Sloan, M. M. (2007). The real self and inauthenticity: The importance of self-concept anchorage for emotional experiences in the workplace. *Social Psychology Quarterly*, *70*(3), 305–318.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159–176.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*(9), 1468–1479.

Strohminger, N., & Nichols, S. (2016). *Psychopathy and the moral self.* (Unpublished manuscript.)

Tajfel, H., & Turner, J. C. (1979/2001). An integrative theory of intergroup conflict. In M. A. Hogg & D. Abrams (Eds.), *Intergroup relations: Essential readings* (pp. 94–109). New York, NY: Psychology Press.

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, *9*(1), 37–43.

Turner, R. H. (1976). The real self: From institution to impulse. *American Journal of Sociology*, *81*(5), 989–1016.

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review*, *4*(3), 387–392.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263.

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*(3), 327–335.