The True Self: A Psychological Concept Distinct From the Self

Nina Strohminger, Joshua Knobe, and George Newman
Yale University

Abstract
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

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 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 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...
Table 1. A Comparison of Key Differences Between the True Self and the Self More Generally

<table>
<thead>
<tr>
<th>The Self</th>
<th>The True Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encompasses entire range of personal features</td>
<td>Emphasizes moral features</td>
</tr>
<tr>
<td>Valence-independent; can be positive or negative</td>
<td>Valence-dependent; positive by default</td>
</tr>
<tr>
<td>Perspective (first- or third-person) dependent</td>
<td>Perspective-independent</td>
</tr>
<tr>
<td>Cross-culturally variable</td>
<td>Cross-culturally stable</td>
</tr>
</tbody>
</table>

Historically, the true self-concept has figured into psychological research in two rather different ways. Because the true self is a commonly held belief among 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 (Bem, 1973; Kernis & Goldman, 2006; Koole & Kuhl, 2003; Maslow, 1943; Masterson, 1988; Rogers, 1961). 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). (Not everyone agrees; Foucault’s charmingly derisive term for Rogers and his ilk is “the Californian cult of the self”; Foucault, 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 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, the mind contributes to personal identity more than the 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 (Prinz & Nichols, 2016; Strohminger & Nichols, 2014). This pattern is quite robust. It shows up regardless of the source of 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 (Arpaly & Schroeder, 1999; Frankfurt, 1988). 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 (Koole & Kuhl, 2003; Masterson, 1988; Rogers, 1961; Sloan, 2007; Turner, 1976), the ideal self (Chodorkoff, 1954; Higgins, 1987; D. T. Kenny, 1956), the authentic self (Cable, Gino, & Staats, 2013; Johnson & Boyd, 1995), the intrinsic self (Arndt, Schimel, Greenberg, & Pyszczynski, 2002; Schimel, Arndt, Pyszczynski, & Greenberg, 2001), 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 (Bargh, McKenna, & Fitzsimons, 2002; Johnson, Robinson, & Mitchell, 2004; Landau et al., 2011; Newman, Lockhart, & Keil, 2010; Schlegel, Hicks, Arndt, & King, 2009; Sheldon, Ryan, Rawsthorne, & Ilardi, 1997). Although 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 to something we see both in our own selves and in other people.
change (the aging process, medical interventions, supernatural events) and regardless of the type of moral feature (disposition, behavior, or belief) (Heiphetz, Strohminger, & Young, 2017; Molouki & Bartels, 2017). Moral traits are considered to be the most deeply rooted, causally central aspect of a person's identity (Chen, Urminsky, & Bartels, 2016), 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 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 than patients with diseases that primarily impact memory (Alzheimer's) or motor function (amyotrophic lateral sclerosis). Even in neurodegenerative diseases that primarily impact 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. Although 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, 2015; Goodwin, Piazza, & Rozin, 2014). This commonality suggests 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 (Molouki & Bartels, 2017; Tobia, 2016). 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, Vess, Schlegel, & Hicks, in press). 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 who someone is (Haslam, Bastian, & Bissell, 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, 2017). 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; Christy, Kim, et al., in press; Molouki & Bartels, 2017; Tobia, 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 (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., in press). To date, the only manipulation that has been able to eliminate this positivity bias 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. Though we are perfectly willing to conceive of other people as bad, we are unwilling to see them as bad deep down.

These findings demonstrate a key respect in which people’s understanding of the true self differs from their understanding of the self. People are happy to attribute dishonorable motives to a person (e.g., S. T. Fiske, 1980; Wojciszke, Brycz, & Borkenau, 1993), but they are reluctant to attribute such motives to the person’s true self. No
matter how heinous the act, there is a strong tendency to conclude that the true self is calling the person to do what is right.

The True Self Is Perspective-Independent

A classic finding 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—rather, 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 someone to the sinner category (Klein & O’Brien, in press; Reeder & Coovert, 1986). 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 (Gilbert & Malone, 1995; Jones & Harris, 1967).

When people judge themselves, this pattern reverses. People routinely overestimate their own knowledge and abilities (Gilovich, 1991; Kruger & Dunning, 1999; Rozenblit & Keil, 2002; West & Stanovich, 1997) 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 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, Gilovich, & Ross, 2004; Pronin, Lin, & Ross, 2002). However, 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, 2017), 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., 2017). 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 irregularity 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., in press). Similarly, Americans and the Japanese believe that negative traits (such as being unkind or having bad vision) will tend to “correct themselves” over time, belying an implicit 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—appear to be culturally invariant. Of course, the present account also predicts a specific kind of cultural difference. When different cultures have different views about what is morally good, these different cultures should show correspondingly different patterns of true self attributions. 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.
 hauls for actions that were driven by overwhelming, irre-

...
Freitas & Cikara, 2016). Although people might continue to acknowledge that the outgroup members differ from themselves on a variety of dimensions, reminding them that outgroup members share 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 judges 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., in press). 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 do not 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 people’s true self attributions. Regardless of culture, perspective, or personality, people tend to see the true self as morally good. Why is it this the case? We see two possible explanations, 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, 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 nonhuman 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 people’s reasoning.

One way to spell out this second possibility is that judgments about the true self are a product of psychological essentialism (Barsalou, 1985; Bloom, 2004, 2010; Gelman, 2003; Keil, 1989; Lynch, Coley, & Medin, 2000; 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 (De Freitas et al., in press; Knobe, Prasada, & Newman, 2013). The propensity to see others as having a good true self might then be explained in terms of this more general way that 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, 2015). Perhaps the same is true of people, with morality being seen as the human telos. Another, nonmutually exclusive, possibility is that moral traits are especially important because they are essential for the maintenance of social bonds (Heiphetz et al., 2017; Strohminger & Nichols, 2015). Because humans are hypersocial creatures, they may be keenly tuned to tracking the moral features of the people around them.

Although 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. The popularity of this idea may be made possible by the notion of the true self.

**Coda: Does the true self exist?**

In this article, 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. First, the true self depends on the values of the observer. If someone thinks homosexual urges are wrong, he or she will say the desire to resist such urges represents the true self (Newman, Knobe, & Bloom, 2014). And if he or she scores high in psychopathy, he or 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.

Second, 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 consistently deplorable are deemed good “deep down.” Think of Anne Frank, hunkered in her secret annex, declaring to her diary, “I still believe, in spite of everything, that people are truly good at heart” (1947/1995, p. 332). 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.

**Acknowledgments**

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.

**Declaration of Conflicting Interests**

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

**References**


Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in cat-


Psychological Concept Distinct From the Self


