Deep down my enemy is good: Thinking about the true self reduces intergroup bias

Julian De Freitas*, & Mina Cikara

Department of Psychology, Harvard University, United States

ARTICLE INFO

ABSTRACT

Intergroup bias — preference for one's in-group relative to out-groups — is one of the most robust phenomena in all of psychology. Here we investigate whether a positive bias that operates at the individual-level, belief in a good true self, may be leveraged to reduce intergroup bias. We find that even stereotypically threatening out-group agents are believed to have a good true self (Experiment 1). More importantly, consideration of an in-group and out-group members' true self reduces intergroup bias, both in the form of explicit evaluative judgments (Experiment 2) and actual donation behavior (Experiment 3). Across studies, the palliative effects of thinking of an individual's true self generalize to that individual's entire group. In sum, a simple intervention — thinking about another's true self — reduces the gap in how people evaluate and treat out-group relative to in-group members. We discuss implications of these findings for conflict reduction strategies.

"He's not a bad guy, deep down," I said. My dad slipped the key into the door. "Deep down, no one is." — Aaron Starmer, The Riverman.

"We're all Muslims deep down. We all yearn for peace." — Boston Police Commissioner William B. Evans, speaking at the Islamic Society of Boston Cultural Center.

Intergroup conflict is one of humanity's greatest challenges. By at least one estimate, over 170 million civilians have perished as casualties of various forms of intergroup violence (Woolf & Hulsizer, 2004). As such, conflict reduction interventions have become a top priority for policy makers and researchers alike (Cohen & Insko, 2008). Psychologists have reported some success reducing prejudice and conflict with a variety of approaches, including: highlighting superordinate goals and identities (Gaertner & Dovidio, 2000), training regulation of negative emotions (Halperin, Porat, Tamir, & Gross, 2013), fostering empathy across groups (Cikara, Bruneau, & Saxe, 2011), and initiating real as well as imagined contact between groups (Crisp & Turner, 2009; Pettigrew & Tropp, 2008). However, research examining the efficacy of these approaches reveals that positive effects may be short-lived (e.g., Bruneau & Saxe, 2012), may not generalize to entire groups (Brewer & Miller; 1984), and may backfire, particularly when parties are of unequal status or power (Dixon, Durrheim, & Tredoux, 2007; Dovidio, Gaertner, & Saguy, 2009; Vorauer & Sasaki, 2009). These 'ironic' effects of intervention come about often by bringing unsavory information into focus (Zaki & Cikara, 2015) or by reinforcing an unjust status quo (Dixon et al., 2010).

All of these interventions start from the recognition that intergroup bias — the preferential evaluation and treatment of in-group relative to out-group members — is a fundamental facet of human psychology (Hewstone, Rubin, & Willis, 2002). This bias manifests across real and arbitrary groups in resource allocation (Tajfel & Turner, 1979), trait evaluations (Locksley, Ortiz, & Hepburn, 1980), implicit bias (Ashburn-Nardo, Voils, & Monteith, 2001), and physiological responses (Cikara & Van Bavel, 2014). In overt conflict, mere in-group preference is combined with out-group hostility, fostering anger, disgust, and aggressive behavior (Cuddy, Fiske, & Glick, 2007; Mackie & Hamilton, 1993). Informed by classic and contemporary theories of intergroup relations, most conflict interventions aim to close these gaps by targeting *group-related* cognitions and emotions (e.g., out-group empathy or anxiety, common identity, familiarity). Here we propose a novel approach in which we fight intractable intergroup bias with another robust, individual-level bias: the good true self bias (De Freitas, Cikara, Grossmann, & Schlegel, 2017; Strohminger, Knobe, & Newman, in press).[1]

1. The good true self and potential links to intergroup behavior

People often explain the behavior of *individuals* by appealing to the concept of a good true self (Newman, Bloom, & Knobe, 2014). An emerging consensus is that this belief in a good true self is a form of psychological essentialism, whereby people tend to view morally good

---

* Corresponding author at: William James Hall, 33 Kirkland St., Cambridge, MA 02138, United States.
  E-mail address: defreitas@g.harvard.edu (J. De Freitas).
[1] We use the term 'bias' to refer to a response tendency, rather than to a normative departure from rationality.

traits as an essential part of a person's identity (De Freitas et al., 2017a; De Freitas, Tobia, Newman, & Knobe, 2016; Strohminger, Knobe, & Newman, in press). In particular, the belief shows at least eight characteristics of psychological essentialism that also make it potentially relevant to intergroup bias.

First, when evaluating a hypothetical agent, people reason as if there is something within the agent calling him or her to behave in a manner that is *morally good*. For example, if the agent changes from behaving badly to behaving virtuously then participants are more likely to report that this improvement reflects the emergence of the agent's true self; conversely, if the agent changes from behaving virtuously to behaving badly, participants report that this deterioration reflects a movement away from the agent's true self (Newman et al., 2014). Since the valence of this belief (the true self is morally *good*) operates in the opposite direction than the negative attitudes typically felt toward out-group members, leveraging the true self bias within the context of intergroup judgments could reduce negative attitudes toward out-group members.

Second, the true self is equated with the fundamental *identity of a person*. Various identity judgments (i.e., whether the person is still the same person) are consistently influenced by the removal of morally good traits more than the removal of morally bad traits or even a host of other mental faculties, including personality, memory, perception, and preferences (Prinz & Nichols, in press; Strohminger & Nichols, 2014; Tobia, 2016). In other words, when the morally good traits are removed, people are inclined to say that the person is no longer the same. If the good true self is believed to constitute the identity of all humans, then it is possible that people will even believe that the fundamental identity of an out-group member is morally good.

Third, people believe that the true self is a *stable, inherent part of a person*. Specifically, people rate personality traits that they deem central to a person's identity as more "innate" and stable over time than other traits (Haslam, Bastian, & Bissett, 2004). Therefore, it is possible that even out-group members are viewed as having an inherent, morally good true self. It might just be that, from the standpoint of the observer, this inherent part of the out-group member is less salient or believed to be suppressed or otherwise not expressed.

Fourth, people believe that there is a *boundary* between the reality of the true self and the appearance of one's 'surface self.' People spontaneously describe the true self as a physical entity "inside" or "beneath the surface" (of the extrinsic self) that can "grow", "expand" or be "expressed" (Bench et al., 2015; Moser, 2007). Since intergroup bias involves a tendency to over-emphasize surface-level features of a person (e.g., their skin tone), thinking about a person's true self may lead one to focus on more stable, invariant aspects of an out-group member rather than these surface-level differences. It is less clear what thinking about the true self of an in-group member might do, since people already have a baseline tendency to view in-group members in an overly optimistic way relative to out-group members (e.g., Hewstone, Rubin, & Willis, 2002; Taylor & Brown, 1994), suggesting that they might already emphasize the morally good characteristics of in-group members. Therefore, one possibility is that thinking about the true self of an in-group member induces even more positive attitudes toward the in-group. Another (somewhat paradoxical) possibility is that thinking about an in-group member's true self leads to more realistic evaluations of the in-group. That is, if people are asked to think about whether an in-group member's behaviors reflect their true versus surface self, then they might be reminded that the in-group member is not uniformly good, but also has a surface self that is not always an expression of the good true self.

Fifth, belief in a good true self is *perspective-independent*; people regard both their own true selves (Bench et al., 2015) and the true selves of others (Bench et al., 2015; Newman, Bloom, & Knobe, 2014) as fundamentally good. This stands in contrast to a large body of work on the self as a whole that shows robust perspective-dependent asymmetries in a variety of domains, such as fundamental attribution error (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Taylor

& Brown, 1994). As such, it is possible that this third-person attribution of a good true self is not limited to non-threatening others, but also extends to out-group members.

Sixth, the good true self bias is *found in two unlikely boundary cases*: 1) in interdependent cultures, where less emphasis is placed on the individual as a separate entity from others, and 2) in misanthropes, who have explicitly negative views about humanity (De Freitas et al., 2017b). The fact that belief in a good true self is robust across these boundary conditions provides support for the hypothesis that this belief is a fundamental aspect of people's commonsense understanding of others, and thus may have widespread consequences for other aspects of cognition. Of relevance to the current studies, it is possible that this same bias is also resilient to intergroup bias. At the same time, it is reasonable to predict that the good true self bias will not apply to out-group members, since aside from the strength of intergroup bias, it is well documented that people tend to think of the essence of an out-group as a negative category, e.g., 'the essence of Arab immigrants' (Haslam, Rothschild, & Ernst, 2000, 2002).

Here we emphasize the seventh relevant feature of the good true self bias, which is that it falls out of reasoning about the essence of an *individual person*. Extending this idea, it may be that when thinking about the essence of an individual out-group member, e.g. 'the essence of Alhadin', or 'the essence of Jafri', the same positive values normally associated with true selves are recruited. If so, a positive bias that falls out of thinking about the essence of an *individual person* could be leveraged to potentially reduce a negative bias that falls out of thinking about the essence of a disliked out-*group*. Such framing effects (Tversky & Kahneman, 1985) are well documented, including within the context of intergroup judgments. For instance, merely changing the framing from "a group of people" to "people in a group" leads to increased mind perception and sympathy for out-group members (Cooley et al., 2017). Similarly, framing out-group members in terms of their true versus surface or group-identified selves might lead to a positive framing effect on intergroup judgments.

Eighth (and finally) the true self is viewed as *diagnostic of an agent's mental states*. For instance, agents are more likely to be judged as happy or strong-willed when they are believed to be expressing their true selves than when they behave in a manner that is believed to conflict with their true selves (Newman, De Freitas, & Knobe, 2015; Phillips, De Freitas, Mott, Gruber, & Knobe, 2017).[2] Furthermore, intuitions about the true self explain these effects over and above other factors, such as the extent to which the agent's behavior is perceived to be in line with the agent's meta-desires (Pizarro, Uhlmann, & Salovey, 2003). These facts are relevant to intergroup bias because individuation interventions that make people focus on the mental states of an out-group member are more effective at reducing intergroup bias (e.g., in empathy; Bruneau, Cikara, & Saxe, 2015) than those that emphasize targets' surface features such as their physical characteristics. Because belief in a good true self both (i) refers to a particular individual, and (ii) is recruited in order to interpret an agent's mental states in particular, framing an intergroup judgment in terms of the true versus surface self might reduce intergroup bias. Specifically, thinking of out-group members in terms of individuals with more nuanced parts (true vs. surface self) could move people away from a polarizing representation of us vs. them.

In sum, the belief in a good true self consists of various features that appear relevant (and we predict resistant) to intergroup bias. Indeed, since belief in a good true self appears to rely on the more fundamental cognitive tendency of psychological essentialism (De Freitas et al., 2017a), it may be that invoking this concept is an especially potent way to reduce intergroup bias.

---

[2] These findings dovetail nicely with empirical work stemming from Self-Determination Theory, indicating that people do not associate their immoral behaviors with their broader core values (i.e., they compartmentalize bad behaviors; Ryan & Deci, 2000).