Journal of Experimental Social Psychology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

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

Keywords:
Intergroup bias
True self
Essentialism
Lay theories

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.

http://dx.doi.org/10.1016/j.jesp.2017.10.006
Received 31 October 2016; Received in revised form 14 October 2017; Accepted 15 October 2017
0022-1031/ © 2017 Elsevier Inc. All rights reserved.

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).

2. Current experiments and hypotheses

Here we test whether even stereotypically threatening out-group members are believed to have good true selves (Experiment 1), and whether there is an asymmetry in attributions of a good true self to in-group and out-group targets as a function of whether the agent improves or deteriorates. If indeed out-group targets are equally likely to be attributed good true selves as in-group targets, we predict that asking participants to consider individual out-group members' true selves *prior* to making group-level evaluative judgments (Experiment 2) and group-level resource allocation decisions (Experiment 3) will reduce intergroup bias (relative to when participants make these judgments and decisions without first considering targets' true selves).

2.1. Experiment 1: examining true self bias for in-group versus out-group members

In this experiment, we test whether two stereotypically threatening out-group members — an Arab and an Arab immigrant in the U.S. — are equally likely as an in-group member — a white U.S. American — to be attributed a good true self. Specifically, we tested whether, for both agents, moral improvements are thought to reflect the true self more than moral deteriorations. Given the dominant phenomenon of intergroup bias, this pattern of results would be a discovery in its own right, suggesting that intergroup bias does not extend to essentialist reasoning about individual members of the out-group.

It is worth noting that the following data were collected approximately two months after the San Bernadino mass shooting in the USA, which was widely attributed to "Islamic terrorism". If historic context had any effect, it was to make our hypothesized results less likely.

2.1.1. Methods

2.1.1.1. Participants and exclusions. We recruited 1020 participants from the United States through Amazon's Mechanical Turk, an online labor crowdsourcing platform (see Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2013; Ipeirotis, 2010; Paolacci, Chandler, & Ipeirotis, 2010; Berinsky, Huber, & Lenz, 2012), paying them 25c each. We aimed to obtain a final target $N = 600$ (100 per condition). This sample size provides 80% power to detect an effect $\eta_p^2 = 0.125$ (effect size is based on those observed in previous research on the true self; see Newman et al., 2014). As planned a priori, we excluded from the analysis participants who incorrectly answered an attention check at the beginning of the study, or two comprehension questions at the end of the study. We also excluded participants who did not self-identify as being of White ethnicity (all questions used for exclusions are provided in the methods below), yielding a final sample of $N = 613$ ($M_{age} = 36.0$, 53.5% female).

2.1.1.2. Materials and procedure. Participants were randomly assigned to one of six conditions in a 2 (Condition: improvement vs. deterioration) × 3 (Ethnic group: Arab native vs. Arab Immigrant vs. White American) between-subjects design. As a robustness check, we used 6 different vignettes describing different agents (e.g. opposing/supporting terrorism, respecting/mistreating minorities, being a teetotaler/alcoholic), presented between-subjects (see link to archived materials and Newman et al., 2014). These vignettes were chosen because they span a variety of moral improvements and deteriorations, and have previously been shown to elicit a reliable good true self bias across cultures (i.e., participants consistently say that moral improvements reflect the agent's true self more than moral deteriorations; De Freitas et al., 2017b).

Participants were first introduced to the agent they would be learning about, "*Imagine an individual named* [*name*]. [*Name*] *is different from you in almost every way—he has a different occupation and prefers different things than you.*" This introduction is typically used in studies of the true self (De Freitas et al., 2017b; Newman, Bloom, & Knobe, 2014) in order to minimize the effect of perceived similarity between the participant and agent on more arbitrary factors (e.g. same job, lifestyle), since people tend to like those who are similar to them on just about any dimension (Ross, Greene, & House, 1977). For the Arab native and Arab immigrant conditions we translated the English names into Arab equivalents, e.g., Jeffrey to Jafri, or Chad to Chahid. The English and Arabian names did not differ on typical linguistic dimensions, including number of syllables, characters, vowels, or consonants (Cohen's ds = 1.07, 0.59, 0.55, and 0.30 respectively).

Since we wanted to test whether people would exhibit a good true self bias for out-group members even after they had just answered standard measures of intergroup bias, we presented these standard intergroup measures first, followed by the true self measures. We assessed participant's evaluations of the agent's group using a number of measures from the intergroup psychology literature (Cikara et al., 2014; Schmid, Al Ramiah, & Hewstone, 2014; Van Bavel, Packer, & Cunningham, 2011). These measures included (i) perceived intergroup threat ("Please rate the extent to which you agree with the following statement: People from White [Arab] backgrounds like Jeffrey [Jafri] threaten the American way of life", 1 = strongly disagree, 5 = neither agree nor disagree, 9 = strongly agree), (ii) an attitude thermometer ("How do you feel about people from White [Arab] backgrounds like Jeffrey [Jafri]?" 1 = cold, 5 = neither warm nor cold, 9 = warm), and (iii) group identification, which consisted of three questions ("I [value/like/feel connected to] people from White [Arab] backgrounds like Jeffrey [Jafri]", 1 = not at all, 9 = very much). We created composite in-group and out-group identification scores by averaging the three items for each condition (American items $\alpha = 0.86$; Arab native items $\alpha = 0.88$, Arab immigrant items $\alpha = 0.90$).

Consistent with previous work on the true self, participants then read that the agent underwent a behavioral change from morally good to morally bad, or from morally bad to morally good (with the direction of moral change counterbalanced between subjects). Since the good true self bias occurs when there is a preferential attribution of moral improvements (vs. moral deteriorations) to the true self, testing both kinds of moral changes allowed us to determine whether this difference is also found for out-group members (or whether it would be moderated by target ethnicity). Here is an example of these conditions:

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

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

Following previous experiments on the good true self bias (Newman et al., 2014), participants then answered a forced-choice question about what caused the agent's behavior: "In your opinion, what aspect of Al's [Alhadin's] personality caused him to be a very caring and involved father [not be a very caring father and not be involved in his children's lives]?" The answer choices were (a) His 'true self' (the deepest, most essential aspect of his being), (b) His 'surface self' (the things that he learned from society or others),' and (c) None of the above. The third option included a space in which participants could explain their choice. A second question asked participants about the agent's behavior in relation to his/her true self, "Now that Al [Alhadin] is a very caring and involved father [is not a very caring father and is not involved in his children's lives], to what extent is he being true to the deepest, most essential aspects of his being?" (1 = not at all, 9 = very much so).

Participants then answered two comprehension questions, "What was the ethnicity of the man in the story?" (the answer options were: Black/African American, Hispanic/Latino, Asian/Pacific Islander, Native American/American Indian, Arab/Middle Eastern, White/

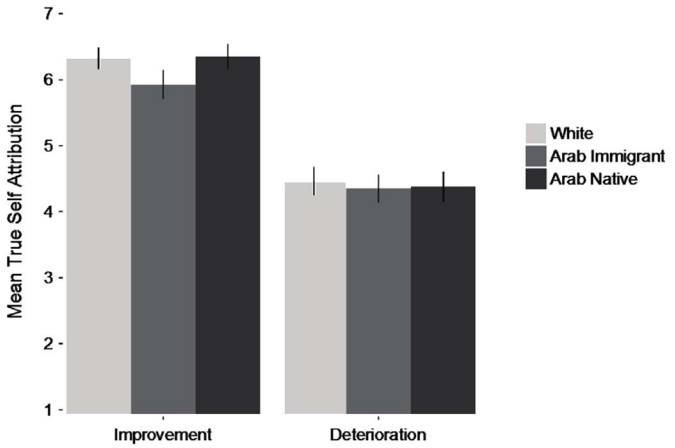

Fig. 1. Mean true self attributions in Experiment 1. Error bars indicate standard error from the mean.

Caucasian, and Other), and "Did the man's behavior improve, deteriorate, or neither (the answer options were: Improve, Deteriorate, Neither). We excluded participants who incorrectly answered these questions based on the condition they were in. Finally, participants reported their ethnicity, age, gender, education, and socioeconomic status, and were debriefed about the purpose of the experiment. We collected educational level and socioeconomic status for the sake of completeness only; we do not report any analyses including these measures.

2.1.2. Results

Where appropriate throughout this paper, we report statistical significance based on the Bonferroni-Holm corrected alpha level, α_{adj}, required to control for multiple comparisons in each multi-factor ANOVA (Cramer et al., 2014).[3] If $p > \alpha_{adj}$, then we conclude that the result is not significant. For all studies, we report all measures, manipulations and exclusions; all sample sizes were determined before any data collection.

2.1.2.1. Intergroup bias. We do not include the condition (improvement vs. deterioration) or vignette factors in these analyses because in Experiment 1 the intergroup bias measures always preceded the vignette in which the agent's behavior improves or deteriorates.

Threat. A 3-way (Ethnicity: White American, Arab native, Arab immigrant) univariate ANOVA revealed the predicted main effect of ethnicity on threat judgments, $F(2, 610) = 6.88$, $p = .001$, $\eta_p^2 = 0.022$. Planned comparisons indicated that the average threat rating for White Americans ($M = 2.59$, $SD = 1.92$) was lower than that for Arab natives ($M = 3.28$, $SD = 2.31$), mean difference = 0.69, 95% CI [0.28, 1.09], $t(411) = 3.29$, $p = .001$, $d = 0.32$, and Arab immigrants ($M = 3.24$, $SD = 2.14$), mean difference = 0.65, 95% CI [0.25, 1.03], $t(417) = 3.23$, $p = .001$, $d = 0.32$.

Attitude thermometer. A 3-way (Ethnicity: White American, Arab native, Arab immigrant) univariate ANOVA revealed the predicted main effect of ethnicity on attitudes, $F(2, 610) = 7.87$, $p = .0004$, $\eta_p^2 = 0.025$. The average attitude thermometer rating for White Americans ($M = 5.67$, $SD = 1.50$) was higher than that for Arab natives ($M = 5.12$, $SD = 1.72$), mean difference = 0.55, 95% CI [−0.86, −0.24], $t(411) = -3.46$, $p = .0006$, $d = -0.34$, and Arab immigrants ($M = 5.12$, $SD = 1.72$), mean difference = 0.55, 95% CI [−0.86, −0.24], $t(417) = -3.51$, $p = .0005$, $d = -0.34$.

Identification. A 3-way (Ethnicity: White American, Arab native, Arab immigrant) univariate ANOVA revealed a main effect of ethnicity on identification, $F(2, 610) = 20.02$, $p = 3.81 * 10^{-9}$, $\eta_p^2 = 0.062$. The average identification rating for White Americans ($M = 5.72$, $SD = 1.59$) was higher than that for Arab natives ($M = 4.89$, $SD = 1.75$), mean difference = 0.83, 95% CI [−1.16, −0.51], $t(411) = -5.05$, $p = 6.7 * 10^{-7}$, $d = -0.50$, and Arab immigrants ($M = 4.76$, $SD = 1.76$), mean difference = 0.96, 95% CI [−1.29, −0.64], $t(417) = -5.90$, $p = 7.68 * 10^{-9}$, $d = -0.58$.

2.1.2.2. Good true self bias. Forced-choice. Following Newman et al. (2014), we recoded the forced-choice item as a binary response with "true self" response as "1" and "surface self" or "other" responses as "0". A logistic regression including only condition found that participants were significantly more likely to attribute moral improvements (59%) than moral deteriorations (30%) to the true self, $z = -7.28$, $p = 3.24 * 10^{-13}$ (see Supplementary Materials Fig. S1). In order to test whether our data favor the hypothesis that there is no difference in good true self attributions to in-group vs. out-group members, we ran a Bayesian logistic regression with an uninformative prior, using the conting package in R (Overstall & King, 2014). The model without an interaction term (condition + ethnicity) was 415.7 times more likely than the model with an interaction term (condition + ethnicity + condition * ethnicity; $BF_{01} = 0.9976/0.0024 = 415.7$), supporting our prediction that the condition effect does not depend on the target's ethnicity.[4]

Scaled-response. As we predicted, a 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab native, Arab immigrant) × 6 (Vignette) univariate ANOVA found only a significant main effect of condition, $F(1, 577) = 112.76$, $p < 2 * 10^{-16} < \alpha_{adj} = 0.0071$, $\eta_p^2 = 0.163$, but no significant main effects of ethnicity, $F(2, 577) = 0.90$, $p = .406$, $\eta_p^2 = 0.003$, or vignette, $F(5, 577) = 1.19$, $p = .314$, $\eta_p^2 = 0.010$, or any higher-order interactions (see Fig. 1). We may not, however, have been sufficiently powered to detect a significant condition × ethnicity interaction. Therefore we also ran a Bayesian ANOVA in JASP (http://www.jasp-stats.org) to compare the relative support for the model with the moderation of condition by ethnic group to a simpler model that included only the main effects of condition and ethnic group. The model that included condition and ethnicity (condition + ethnicity) was 20.0 times more likely than the model that also included a condition × ethnicity interaction (condition × ethnicity; $BF_{01} = 1.00/0.05 = 20.0$).

Collapsing across vignettes and ethnicities, the average true self attribution was higher in the improvement ($M = 6.21$, $SD = 1.93$) than the deterioration condition ($M = 4.40$, $SD = 2.23$), mean difference = 1.81, 95% CI [−2.14, −1.48], $t(611) = -10.73$, $p < 2.2 * 10^{-16}$, $d = -0.87$. In sum, the data favor our prediction that good true self bias is not moderated by target social group.

2.1.3. Discussion

Although participants exhibited the predicted intergroup bias on threat, attitude, and identification measures, even stereotypically threatening out-group members were equally subject to the good true self bias. In other words, in-group and threatening out-group members were judged as having equally good true selves. This effect was the same across groups irrespective of whether the agents improved (in which case their new behavior was viewed as more reflective of the true self) or deteriorated (in which case their new behavior was viewed as

[3] The Bonferroni-Holm correction (aka sequential Bonferroni procedure) sorts all significant p-values of an ANOVA from the smallest to largest p-value, then computes adjusted α levels, α_{adj}, for each of these p-values. For the smallest p-value, α_{adj} equals α divided by the number of tests conducted in the ANOVA; for the next smallest p-value, α_{adj} equals α divided by the number of tests minus 1; and so forth until the largest significant p-value. The significance of each p-value is then evaluated against its respective adjusted alpha. Unlike many other correction procedures, this procedure is not only strict but is also optimized for avoiding *both* Type 1 and 2 errors (Cramer et al., 2014).

[4] For more information on Bayes factors and priors in psychological research, see Wagenmakers et al. (2016, in press).

less reflective of the true self).

2.2. Experiment 2: reducing intergroup bias in explicit evaluative judgments

Experiment 2 used the same methods as Experiment 1, but also counterbalanced the order in which the intergroup judgments and true self measures were presented. Furthermore, since no significant difference emerged for judgments toward Arab natives versus Arab immigrants in Experiment 1, we used only the group that received the most negative judgments overall in that experiment: Arab immigrants. We predicted that participants who considered either in-group or out-group members' true selves prior to (as compared to after) making evaluative group judgments would exhibit reduced intergroup bias (i.e., relatively more neutral evaluations of both the in-group and out-group).

Notice that in principle participants could be recruiting the notion of a true self when reasoning not only about moral improvements but also moral deteriorations, since for both conditions participants judge whether or not the behavior reflects the agent's true self. Yet given that the true self judgment task only asks participants to rate the extent to which a given behavioral change reflects the true self (and finds preferential attribution of moral improvements), this measure does not allow us to determine whether people are also activating the concept in the deterioration condition. If they are, then we should expect this activated concept to go on to reduce intergroup bias in *both* moral improvement and deterioration conditions (e.g., “he is simply acting in accordance with his true self in the improvement condition, or not in the deterioration condition”). In contrast, if the concept is only activated when reasoning about moral improvements, we should only expect to see a reduction in intergroup bias in the improvement condition.

2.2.1. Methods

2.2.1.1. Participants and exclusions. We recruited 1327 participants from the United States through Amazon's Mechanical Turk, paying them 25c each. We aimed for the same sample size per condition as in Experiment 1 ($N = 100$ after exclusions). As planned a priori, we excluded from the analysis participants who took part in Experiment 1, or who incorrectly answered the attention check at the beginning of the study or two comprehension questions at the end of the study. We also excluded participants who did not self-identify as being of White ethnicity, yielding a final sample of $N = 759$ ($M_{age} = 37.1$, 56.6% female).

2.2.1.2. Materials and procedure. Participants were randomly assigned to one of eight conditions in a 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab Immigrant) × 2 (Order: True Self Judgments After, Before Intergroup Judgments) between-subjects design. As a robustness check, we again used 6 different vignettes describing different agents, presented between-subjects (taken from Newman et al., 2014).

As in Experiment 2, participants were first introduced to the agent they would be learning about. They then received either the intergroup judgment measures followed by the true self measures (including the accompanying vignettes), or vice versa. Finally, they answered the same comprehension and demographic items as used in Experiment 1.

2.2.2. Results

2.2.2.1. Good true self bias. First we examined whether we replicated the main finding of Experiment 1: that participants were no less likely to attribute a good (i.e., improved) true self to a threatening out-group member than to an in-group member.

Forced-choice. The forced-choice responses were recoded as in Experiment 1.

A logistic regression including only condition found that participants were significantly more likely to attribute moral improvements (58%) than moral deteriorations (29%) to the true self, $z = -7.77$, $p = 7.96 * 10^{-15}$ (see Supplementary Materials Fig. S2). To test whether the data favor the hypothesis that there is no difference in good true self attributions to in-group vs. out-group members, we ran a Bayesian logistic regression. A model without an interaction term (condition + ethnicity + order) was the most likely of all models, and 3.8 times more likely than the next most supported model with an interaction term (condition + ethnicity + order + condition * ethnicity; $BF_{01} = 0.7104/0.1864 = 3.8$), supporting our hypothesis.

Scaled item. A 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab immigrant) × 2 (Order: True Self Judgments After, Before) × 6 (Vignette) univariate ANOVA found only a significant main effect of condition, $F(1, 711) = 201.55$, $p < 2 * 10^{-16} < \alpha_{adj} = 0.0033$, $\eta_p^2 = 0.219$, and a marginal main effect of vignette, $F(5, 711) = 3.59$, $p = .0032 < \alpha_{adj} = 0.0036$, $\eta_p^2 = 0.024$. All vignettes elicited higher true self ratings for the improvement than deterioration condition, but some elicited larger effects than others (see Supplementary Materials Fig. S3). There were no main effects of ethnicity, $F(1, 711) = 4.86$, $p = .028 > \alpha_{adj} = 0.0038$, $\eta_p^2 = 0.007$, or order, $F(1, 711) = 0.58$, $p = .445$, $\eta_p^2 = 0.0007$. We also did not find a significant ethnicity × order interaction, $F(1, 711) = 2.90$, $p = .089$, $\eta_p^2 = 0.004$, a condition × ethnicity interaction $F(1, 711) = 1.95$, $p = .163$, $\eta_p^2 = 0.003$, or any significant 3-way or 4-way interactions (all $\eta_p^2 \leq 0.01$). In order to test whether the data favor a model in which good true self attributions do not depend on the target's ethnicity, we also ran a Bayesian ANOVA. The model that included condition and vignette (condition + vignette) received the greatest support of all models, and was 4.2 times more likely than the next most supported model that included an interaction term (vignette + condition + ethnicity + condition * ethnicity), $BF_{01} = 1.00/0.24 = 4.2$.

Collapsing across vignettes and ethnicities, the average true self attribution was higher in the improvement ($M = 6.51$, $SD = 1.89$) than the deterioration condition ($M = 4.32$, $SD = 2.38$), mean difference = 2.19, 95% CI [−2.50, −1.89], $t(757) = -14.09$, $p < 2.2 * 10^{-16}$, $d = -1.02$. Therefore, mirroring the results from the forced-choice measure, we once again replicated the good true self bias for judgments of both in-group and out-group members.

2.2.2.2. Intergroup bias. Threat. A 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab immigrant) × 2 (Order: True Self Judgments After, Before) × 6 (Vignette) univariate ANOVA revealed significant main effects of order, $F(1, 711) = 14.03$, $p = .0036 < \alpha_{adj} = 0.004$, $\eta_p^2 = 0.018$, and ethnicity, $F(1, 711) = 10.36$, $p = .001 < \alpha_{adj} = 0.0038$, $\eta_p^2 = 0.013$. These main effects were qualified by the predicted significant ethnicity × order interaction, $F(1, 711) = 19.87$, $p = 9.36 * 10^{-6} < \alpha_{adj} = 0.0033$, $\eta_p^2 = 0.028$. There was no significant main effect of vignette, $F(5, 711) = 3.30$, $p = .006 > \alpha_{adj} = 0.0042$, $\eta_p^2 = 0.024$, nor a significant vignette × order interaction, $F(5, 711) = 2.35$, $p = .039 > \alpha_{adj} = 0.0045$, $\eta_p^2 = 0.015$. Finally, there was also no significant main effect of condition, condition × order interaction, or 3-way or 4-way interactions (all $\eta_p^2 \leq 0.01$).

In order to unpack the significant ethnicity × order interaction, we conducted simple contrasts on least-squares means using the lsmeans package in R (Lenth, 2016), which maintains experiment-wide error. First, we looked at those cases in which the intergroup judgment measures were presented *prior* to the true self judgments (replicating the design of Experiment 1). The average threat rating for White Americans ($M = 2.37$, $SE = 0.14$) was significantly lower than that for Arab immigrants ($M = 3.53$, $SE = 0.17$), mean difference = 1.16, 95% CI [0.58, 1.73], $t(711) = 5.19$, $p < .0001$, $d = 0.53$. When the true self measures were presented first, however, the average threat rating for White Americans ($M = 3.54$, $SE = 0.15$) did not differ significantly from that for Arab immigrants ($M = 3.29$, $SE = 0.17$), mean difference = −0.25, 95% CI [−0.82, 0.33], $t(711) = 1.09$, $p = .698$, $d = 0.11$. See Fig. 2 for a summary of results.

Attitude thermometer. A 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab immigrant) × 2

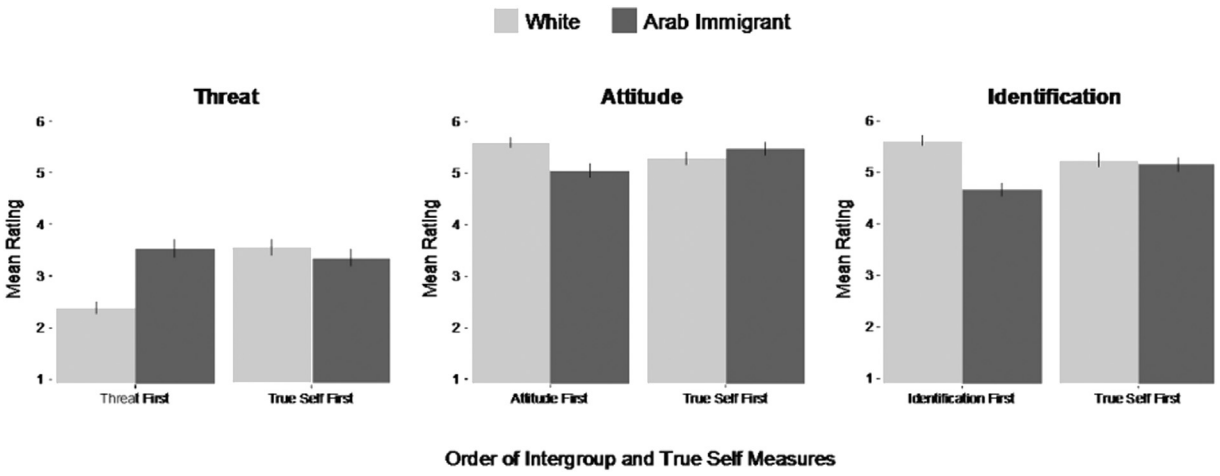

Fig. 2. Mean intergroup ratings for threat, attitude, and identification, in Experiment 2. Error bars indicate standard error from the mean.

(Order: True Self Judgments After, Before) × 6 (Vignette) univariate ANOVA revealed only the predicted, but marginally significant ethnicity × order interaction, $F(1, 711) = 8.09$, $p = .0046 > \alpha_{adj} = 0.0033$, $\eta_p^2 = 0.012$. Note that technically using an adjusted alpha here is overly conservative, since there was nothing exploratory about this predicted interaction (Cramer et al., 2014), and without this correction the interaction easily meets standard criteria for significance. Nonetheless, we report corrected alphas here for consistency.

None of the other tests was significant. There was no significant main effect of vignette, $F(5, 711) = 2.31$, $p = .043 > \alpha_{adj} = 0.0038$, $\eta_p^2 = 0.016$, nor significant main effects of ethnicity, condition, condition × order interaction, or 3-way or 4-way interactions (ethnicity $p = .137$; all other $ps > 0.250$). There was also no significant vignette × ethnicity × condition interaction, $F(5, 711) = 2.61$, $p = .024 > \alpha_{adj} = 0.0036$, $\eta_p^2 = 0.018$.

In order to unpack the ethnicity × order interaction, we again conducted simple contrasts using lsmeans. First, we looked at those cases in which intergroup judgment measures were presented before true self judgments. The average attitude thermometer rating for White Americans ($M = 5.59$, $SE = 0.12$) was significantly higher than that for Arab immigrants ($M = 5.02$, $SE = 0.14$), mean difference = 0.57, 95% CI [−1.04, −0.10], $t(711) = 3.12$, $p = .010$, $d = 0.32$. When the true self measures were presented first, however, the average attitude thermometer rating for White Americans ($M = 5.30$, $SE = 0.12$) did not differ significantly from that for Arab immigrants ($M = 5.47$, $SE = 0.14$), mean difference = −0.17, 95% CI [−0.30, 0.64], $t(711) = 0.93$, $p = .789$, $d = 0.10$. See Fig. 2 for a summary of results.

Identification. As in Experiment 1, we first created composite in-group and out-group identification scores by averaging the three identification questions for each condition (American items $\alpha = 0.93$; Arab immigrant items $\alpha = 0.89$). A 2 (Condition: Deterioration, Improvement) × 2 (Ethnicity: White American, Arab immigrant) × 2 (Order: True Self Judgments After, Before) × 6 (Vignette) univariate ANOVA revealed a significant main effect of ethnicity, $F(1, 711) = 16.13$, $p = 6.55 * 10^{-5} < \alpha_{adj} = 0.0033$, $\eta_p^2 = 0.022$, and, critically, the predicted ethnicity × order interaction, $F(1, 711) = 10.55$, $p = .001 < \alpha_{adj} = 0.0036$, $\eta_p^2 = 0.015$. There was no main effect of condition, significant condition × order interaction, or 3-way or 4-way interactions (vignette × ethnicity × condition $p = .062$; all other $ps > .250$).

In order to unpack the ethnicity × order interaction, we again conducted simple contrasts using lsmeans. We first examined just those cases when the intergroup judgment measures were presented first. The average identification rating for White Americans ($M = 5.61$, $SE = 0.12$) was significantly higher than that for Arab immigrants ($M = 4.63$, $SE = 0.14$), mean difference = 0.98, 95% CI [−0.97, 0.19], $t(711) = 5.17$, $p < .0001$, $d = 0.53$. When the true self measures were presented first, however, the average identification rating for White Americans ($M = 5.25$, $SE = 0.13$) did not differ significantly from that for Arab immigrants ($M = 5.17$, $SE = 0.14$), mean difference = 0.08, 95% CI [−0.57, 0.40], $t(711) = 0.43$, $p = .973$, $d = 0.05$. See Fig. 2 for a summary of results.

2.2.3. Discussion

The good true self bias replicated for both in-group and out-group targets: participants judged improvements as a better reflection of targets' true selves than deteriorations. More importantly, thinking first about the true self for in-group and out-group members (irrespective of whether it was a deterioration or improvement vignette) significantly reduced intergroup bias in the form of explicit evaluations of threat, attitudes, and identification. Further, note that the intergroup bias was attenuated by both reducing positive evaluations of the in-group and increasing positive-evaluations of the out-group. We did not predict this effect. It may be that thinking of an in-group member in terms of a true versus surface self-reminded participants that in-group members are not uniformly positive, but that in-group members also have surface selves that are not always an expression of their good true selves. Moreover, this finding is not predicted by a pure individuating effect, which should have also boosted (or at least left unaffected) attitudes toward the in-group. Thus, it appears that thinking about the true self has an overall equilibrating effect, simultaneously reducing overly positive attitudes toward us and attenuating overly negative attitudes toward them. This dual effect appears to make it particularly effective at reducing intergroup bias.

2.3. Experiment 3: reducing intergroup bias in donation behavior

Experiment 2 demonstrated that considering in-group and out-group members' true selves reduced intergroup bias in evaluation judgments. However, negative appraisals — both explicit attitudes and implicit associations — are not always highly correlated with discriminatory behavior (Ajzen & Cote, 2008; Talaska, Fiske, & Chaiken, 2008). Therefore in Experiment 3 we changed the outcome variable of interest to an incentive compatible behavioral measure. Participants had to decide how much of a bonus payment they wanted to allocate between an in-group charity versus an out-group charity. Aside from being a behavioral measure, this is an especially strong test of intergroup bias, since any amount given to the out-group charity is also an amount *not* given to the in-group charity. One side effect of this stronger measure, however, is that it removes our ability to directly measure

whether a given change in donation behavior is driven by a reduction in in-group preference, an enhancement in out-group preference, or both. Yet, given the results of Experiment 2 (in which we found evidence for both), we expected that reasoning about *either* an in-group or out-group agent prior to making a donation would lead to a larger proportion of donations given to the out-group charity. Such a result would still be consistent with the interpretation that attitudes toward *both* in-group and out-group agents are shifted after thinking about the true self.

Because we did not find an effect of condition (improvement vs. deterioration) on the attenuation of the threat, attitude, and identification bias in Experiment 2, we include only improvement scenarios here. In sum, for Experiment 3 we predicted only a main effect of order (true self before vs. after donation measure) on donation behavior.

2.3.1. Methods

2.3.1.1. Participants and exclusions. We recruited 678 participants from the United States through Amazon's Mechanical Turk, paying them 25c each. We aimed for the same sample size per cell as in Experiments 1 and 2 ($N = 100$ after exclusions). As planned a priori, we excluded from the analysis participants who took part in Experiments 1 and 2, and who incorrectly answered the attention check at the beginning of the study or two comprehension questions at the end of the study. We also excluded participants who did not self-identify as being of White ethnicity, yielding a final sample of $N = 374$ ($M_{age} = 36.7$, 52.0% female).

2.3.1.2. Materials and procedure. Participants were randomly assigned to one of four conditions in a 2 (Ethnicity: White American, Arab Immigrant) × 2 (Order: Charity Donation First, True Self First) between-subjects design. As a robustness check, we again used 6 different vignettes describing different agents, presented between-subjects (taken from Newman et al., 2014).

For the donation measure, participants were asked how much of a $0.50 bonus they would like to allocate between an in-group charity (American Red Cross) and an out-group charity (Syrian Arab Red Crescent), both of which truly exist:

"We are giving you a 50c bonus that we would like you to split between the following two charities that are devoted to providing humanitarian relief during emergencies: the "American Red Cross" and the "Syrian Arab Red Crescent". We will make this donation on your behalf.

How would you like to split the donation?" (the options were 100% to: American Red Cross, 90% to: ARC, 80% to: ARC, 70% to: ARC, 60% to: ARC, 50% to: EACH, 60% to SARC, 70% to: SARC, 80% to: SARC, 90% to: SARC, 100% to: Syrian Arab Red Crescent). For the purposes of analysis we recoded these choices from 1 (100% to in-group charity) to 11 (100% to out-group charity).

Participants either saw the donation measure first followed by the true self measure (including a vignette), or vice versa. Finally, they answered the same comprehension and demographic items as used in Experiments 1 and 2. On behalf of the participants, we summed all of their donations and donated them to the intended charities.

2.3.2. Results

2.3.2.1. True self question. Forced-choice. The forced-choice responses were coded as in Experiments 1 and 2. A logistic regression including only ethnicity found no significant difference depending on whether participants were judging a White American (57%) vs. an Arab immigrant (61%), $z = -0.71$, $p = .478$. Once again we ran a Bayesian Logistic regression to test whether the data favored a model without (as compared to with) main effects of order and target ethnicity, as well as the interaction between the two. Despite running a huge number of Markov Chain Monte Carlo (MCMC) samples (6,105,000), the regression was unable to find evidence for any model that included either main effect or an interaction. This indicates that the null model with only the intercept is most probable — there was only one posterior model probability with value = 1.[5] Thus, participants did not judge improvements as more indicative of in-group true selves relative to out-group true selves, and it did not matter whether they completed this item before or after the allocation measure.

Scaled response. In line with the forced-choice response results, a 2 (Ethnicity: White American, Arab Immigrant) × 2 (Order: Donation First, True Self First) × 6 (Vignette) univariate ANOVA found no significant main effects of ethnicity, $F(1, 350) = 0.03$, $p = .870$, $\eta_p^2 = 2.78 * 10^{-5}$, order, $F(1, 350) = 0.00$, $p = .989$, $\eta_p^2 = 1.87 * 10^{-10}$, or vignette, $F(5, 350) = 1.54$, $p = .176$, $\eta_p^2 = 2.26 * 10^{-2}$, and, critically, no significant ethnicity × order interaction, $F(1, 350) = 1.64$, $p = .201$, $\eta_p^2 = 0.003$. In the Bayesian ANOVA, the model with only the intercept performed best; Bayes factors for alternative models that included an effect of or an interaction with ethnicity or order all had $BF_{10} < 1$, when compared to the null model with only the intercept.

2.3.2.2. Donation behavior. A 2 (Ethnicity: White American, Arab immigrant) × 2 (Order: Donation First, True self first) × 6 (Vignette) univariate ANOVA revealed only the predicted significant main effect of order, $F(1, 350) = 12.68$, $p = .0004 < \alpha_{adj} = 0.0071$, $\eta_p^2 = 0.038$, conceptually replicating the results of Experiment 2. Collapsing across vignettes, the average donation amount to the Syrian Arab Red Crescent was significantly higher when the true self question was presented first ($M = 4.89$, $SD = 2.87$) than when the donation measure was presented first ($M = 3.81$, $SD = 2.87$), mean difference = 1.08, 95% CI [0.49, 1.66], $t(373) = 3.64$, $p = .0003$, $d = 0.38$ (See Fig. 3).

2.3.3. Discussion

As predicted, and in line with the results of Experiments 2, thinking about the true self of in-group and out-group members increased actual donations to an out-group charity. It is worth noting, however, that in no case did out-group donations exceed in-group donations. The good true self bias in both the in-group and out-group conditions merely had the effect of attenuating intergroup bias in allocation.

3. General discussion

We investigated whether beliefs in a good true self may be leveraged to reduce intergroup bias. Surprisingly, despite the strength and pervasiveness of intergroup bias, people reported that threatening out-group members have equally good true selves as do in-group members (Study 1). Furthermore, instructing participants to consider in-group and out-group true selves first reduced subsequent intergroup bias, both in the form of explicit evaluative judgments (Study 2), and actual donation behavior (Study 3). There are a number of aspects of this effect that make it unique.

To our knowledge, belief in a good true self is one of very few social attributes that does not appear to be impacted by intergroup relations: it did not matter whether people were reasoning about in-group or threatening out-group agents, they consistently believed that the agent was good deep down. The intervention also does not originate at the group level. That is, we found that a positivity bias that operates at the level of the *individual* reduces a bias that falls out of reasoning about individuals in terms of their *group membership*. This may be because thinking about whether an agent's behavior reflected their true vs. surface self-led to a framing effect (Cooley et al., 2017; Tversky and Kahneman, 1985), leading to more nuanced representations of a group member (i.e., true self is good, surface self is bad) rather than thinking of them merely in terms of us (good) vs. them (bad). Although these findings were predicted, they are also somewhat counterintuitive,

[5] For more information on MCMC sampling in psychological research, see Lee and Wagenmakers (2013), Kruschke (2014), and van Ravenzwaaij, Cassey, and Brown (2016).

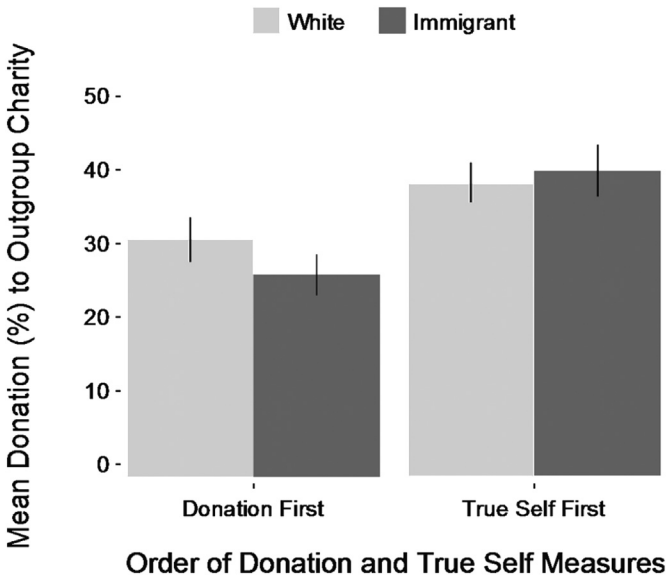

Fig. 3. Mean donations to the out-group charity (Syrian Arab Red Crescent) in Experiment 3. Error bars indicate standard error from the mean.

suggesting that an effective intervention for reducing intergroup bias need not operate at the group level.

Furthermore, the *way* in which the good true self bias reduced intergroup bias was also somewhat surprising. Effective interventions typically aim to reduce negative attitudes toward out-groups specifically (Cohen & Insko, 2008). Yet we found evidence that the good true self bias not only reduced negative attitudes toward out-group members, but also attenuated positive attitudes toward in-group members, leading to net equilibration. Although we did not predict this result, it could be explained by the idea that thinking about the true self led to a more nuanced valence representation of *both* out-group and in-group members (true self is good, surface self is sometimes bad), reminding participants that in-group members are not as uniformly positive as is typically assumed (e.g., Taylor & Brown, 1994). This might be an ironic effect of thinking about in-group members in terms of a true versus surface self: people are reminded that even in-group members are layered — yes, they are good, but they also have superficial aspects. The reverse would happen when thinking about the true self of an out-group member — they are not completely bad, since they are still good deep down. Thus, our findings suggest that intergroup interventions have legroom to operate within *both* out-group and in-group representations, making both more neutral, and therefore more equitable. This effect might fall out of an integral characteristic of the belief in a good true self: the belief that there is an intuitive boundary between the true self (reality) and the surface self (appearance). In contrast, this finding is not explained by a pure individuating effect, which would have improved attitudes toward the in-group or at least left these attitudes unaffected.

More broadly, we believe that the various effects observed here fall directly out of the concept of a good true self, which is thought to arise from *psychological essentialism,* the tendency to understand entities in terms of deeper, unobservable essences (Ahn et al., 2001; Gelman, 2003; Dar-Nimrod & Heine, 2011; Keil, 1989; Medin & Ortony, 1989; Xu, & Rhemtulla, 2005). Just as people show a robust, cross-cultural tendency to posit an unobservable essence for a variety of entities (e.g., Atran, 1993; Brown, 1991; Gil-White, 2001; Hirschfeld, 1998; Sousa, Atran, & Medin, 2002), they appear to posit an unobservable essence of the self, yielding the notion of a 'true self' (for reviews, see De Freitas et al., 2017a; Strohminger, Knobe, & Newman, in press). Moreover, they view the true self as consisting of traits that they morally value (Newman, Bloom, & Knobe, 2014; De Freitas et al., 2017b), yielding the notion of a *good* true self. The belief also has various other characteristics that make it relevant to an intergroup context, including the belief that the true self is a stable, inherent part of a person that is bounded from other parts of the self; the fact that the belief is perspective-independent; and the fact that the true self falls out of reasoning about the essence of an individual person, and is also believed to be diagnostic of a person's mental states. We think that these various features of the concept explain the nuanced effects we report here, and also complement previous work showing that individuation manipulations which emphasize mental states are more effective at reducing intergroup bias than surface features are (Bruneau, Cikara, & Saxe, 2015). Finally, since the belief has already been found to be resilient across boundary cases like interdependent cultures and individual differences in misanthropy, we expected that it might generalize to intergroup reasoning as well.

We also think the essentialist framework provides a more parsimonious account of these effects relative to various other theories. For instance, the results are not consistent with the theory that thinking about good things in general reduces intergroup bias, since making true self judgments about *both* improvements and deteriorations reduced intergroup bias. Furthermore, it is unlikely that thinking about the true self leads one to think about feelings more than behaviors, thereby leading to more empathy (Andersen & Ross, 1984), since the good true self bias has previously been shown to apply to both feelings and overt behaviors (Newman, Bloom, & Knobe, 2014). In contrast to these alternatives, the results suggest that thinking about whether in-group and out-group members' behaviors reflect their true self vs. surface self leads to a more nuanced representation of these individuals, enhancing positive views of an out-group member (which may be overly-negative to start with) and dampening positive views of an in-group member (which may be overly-positive to start with). Future studies can certainly probe this mechanism further, but the main objective of the current studies has been to discover that 1) the good true self also applies to out-group members, and 2) it can thus be cheaply leveraged to reduce intergroup bias in both evaluations and behavior.

Critically, the intervention reported here is incredibly simple; it does not require the application of external, expensive, or elaborate methods and regimens. Instead, it would appear that the main ingredient for change comes conveniently pre-packaged in people's intuitive psychology: thinking of individuals in terms of morally good essences (De Freitas et al., 2016; Strohminger & Nichols, 2014). As such, it may be that this manipulation is particularly potent. Of course, not all essentializing is good, as in the well-documented tendencies to think of the essence of an out-group as a category, e.g. 'the essence of Arab immigrants' (Haslam, Rothschild, & Ernst, 2000, 2002). In contrast, our studies show that asking people to think about the essence of an individual person, e.g. 'the essence of Alhadin', or 'the essence of Jafri', has immediate, powerful consequences for intergroup cognition and behavior.

Our approach is similar to other interventions that activate beliefs or mindsets which do not directly reference the relationship between groups but nevertheless have effects on intergroup dynamics. For example, inducing malleable beliefs about groups (e.g., that groups, in general, can change their basic characteristics) leads Israeli Jews and Palestinian citizens of Israel to express more positive attitudes toward the out-group, and to express a greater willingness to compromise with their out-groups (Halperin, Russell, Trzesniewski, Gross, & Dweck, 2011). Note that these malleable versus fixed prompts did not reference Israel, Palestine, or even the concept of out-groups, though the outcome measures referred specifically to the Israeli-Palestinian conflict (Halperin et al., 2011). Similar to these approaches, the true self bias may avoid the defensive reactions that are sometimes associated with directing participants to adopt out-group perspectives or to engage in a structured interaction, because participants in the current experiments were never directly told that the true selves of in-group and out-group members were essentially good. Instead, they were simply asked to reflect on the individual features of an agent (i.e., on whether a given

behavior was a manifestation of the agent's true self), and the design was entirely between-subjects. As such, participants presumably did not need to compare the out-group members to their own self or to in-group members, which has sometimes been found to induce distinctiveness threats that would likely undermine intergroup interventions (Jetten & Spears, 2003). Future studies should directly test whether invoking a direct comparison between an out-group and in-group target undermines the effect documented here.

In sum, although we tend to view in-group members as mostly good and out-group members as mostly bad, neither is viewed as rotten to the core. Directing people to consider this common, essential good may lead to more nuanced representations of us and them, fostering more equitable treatment across the group divide.

Open practices

The research in this article earned Open Materials and Open Data badges for transparent practices. Experiment materials, data, and analysis code for all three experiments in the paper can be downloaded at: https://osf.io/g6x7f/?view_only=a780cde9131745bebce281fcd0029a9c.

Acknowledgments

We thank Alexander Ly for advising us on the Bayesian analyses and write-up.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2017.10.006.

References

Ahn, W., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., ... Shafto, P. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82*, 59–69.

Ajzen, I., & Cote, N. G. (2008). Attitudes and the prediction of behavior. In W. D. Crano, & R. Preslin (Eds.). *Attitudes and attitude change*. New York: Psychology Press.

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

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology, 81*, 789–799.

Atran, S. (1993). *Cognitive foundations of natural history: Towards an anthropology of science.* Cambridge University Press.

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

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*, 169–185.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*, 351–368.

Brewer, M. B., & Miller, N. (Eds.). (1984). *Groups in contact: The psychology of desegregation*. Academic Press.

Brown, D. E. (1991). *Human universals.* New York: McGraw-Hill118.

Bruneau, E. G., Cikara, M., & Saxe, R. (2015). Minding the gap: Narrative descriptions about mental states attenuate parochial empathy. *PLoS One, 10*(10), e0140838.

Bruneau, E. G., & Saxe, R. (2012). The power of being heard: The benefits of "perspective-giving" in the context of intergroup conflict. *Journal of Experimental Social Psychology, 48*, 855–866.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them intergroup failures of empathy. *Current Directions in Psychological Science, 20*, 149–153.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology, 55*, 110–125.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations an integrative review. *Perspectives on Psychological Science, 9*, 245–274.

Cohen, T. R., & Insko, C. A. (2008). War and peace: Possible approaches to reducing intergroup conflict. *Perspectives on Psychological Science, 3*, 87–93.

Cooley, E., Payne, B. K., Cipolli, W., III, Cameron, C. D., Berger, A., & Gray, K. (2017). The paradox of group mind: "People in a group" have more mind than "a group of people". *Journal of Experimental Psychology: General, 146*(5), 691–699.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... Wagenmakers, E. J. (2014). Hidden multiplicity in multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*, 640–647.

Crisp, R. J., & Turner, R. N. (2009). Can imagined interactions produce positive perceptions?: Reducing prejudice through simulated social contact. *American Psychologist, 64*, 231–240.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*, 631–648.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: On the deceptive determinism of DNA. *Psychological Bulletin, 137*, 800–818.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences*.

De Freitas, J., Newman, G. E., Sarkissian, H., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2017). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*.

De Freitas, J., Tobia, K., Newman, G. E., & Knobe, J. (2016). Normative judgments and individual essence. *Cognitive Science,* 1–21.

Dixon, J., Durrheim, K., & Tredoux, C. (2007). Intergroup contact and attitudes toward the principle and practice of racial equality. *Psychological Science, 18*, 867–872.

Dixon, J., Tropp, L. R., Durrheim, K., & Tredoux, C. (2010). "Let them eat harmony": Prejudice-reduction strategies and attitudes of historically disadvantaged groups. *Current Directions in Psychological Science, 19*, 76–80.

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2009). Commonality and the complexity of "we": Social attitudes and social change. *Personality and Social Psychology Review, 13*, 3–20.

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: The common ingroup identity model.* Philadelphia: Psychology Press.

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

Gil-White, F. J. (2001). Sorting is not categorization: A critique of the claim that Brazilians have fuzzy racial categories. *Journal of Cognition and Culture, 1*, 219–249.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224.

Halperin, E., Porat, R., Tamir, M., & Gross, J. J. (2013). Can emotion regulation change political attitudes in intractable conflicts? From the laboratory to the field. *Psychological Science, 24*, 106–111.

Halperin, E., Russell, A. G., Trzesniewski, K. H., Gross, J. J., & Dweck, C. S. (2011). Promoting the Middle East peace process by changing beliefs about group malleability. *Science, 333*, 1767–1769.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin, 30*, 1661–1673.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39*, 113–127.

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? *British Journal of Social Psychology, 41*, 87–100.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology, 53*, 575–604.

Hirschfeld, L. A. (1998). *Race in the making: Cognition, culture, and the child's construction of human kinds.* MIT Press.

Ipeirotis, P. (2010, March 9). The new demographics of Mechanical Turk. Retrieved from http://www.behind-the-enemy-lines.com/2010/03/newdemographics-of-mechanical-turk.html.

Jetten, J., & Spears, R. (2003). The divisive potential of differences and similarities: The role of intergroup distinctiveness in intergroup differentiation. In W. Stroebe, & M. Hewstone (Vol. Eds.), *European review of social psychology. Vol. 14. European review of social psychology* (pp. 203–241). Hove, UK: Psychology Press.

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

Kruschke, J. (2014). Doing Bayesian data analysis. *A tutorial with R: JAGS, and Stan.* Elsevier: Science.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*, 1–33.

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*, 773–783.

Mackie, D., & Hamilton, D. (1993). *Affect, cognition, and stereotyping.* San Diego, CA: Academic.

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

Moser, K. S. (2007). Metaphors as symbolic environment of the self: How self-knowledge is expressed verbally. *Current Research in Social Psychology, 12*, 151–178.

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

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*, 96–125.

Overstall, A. M., & King, R. (2014). conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software, 58*, 1–27. http://www. jstatsoft.org/v58/i07.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making, 5*, 411–419.

ARTICLE IN PRESS

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology, 38*, 922–934.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General, 146*, 165–181.

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

Prinz, J., & Nichols, S. (2017). Diachronic identity and the moral self. In J. Kiverstein (Ed.). *Handbook of the social mind*London: Routledge (in press).

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

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

Schmid, K., Al Ramiah, A., & Hewstone, M. (2014). Neighborhood ethnic diversity and trust the role of intergroup contact and perceived threat. *Psychological Science,* 1–10.

Sousa, P., Atran, S., & Medin, D. (2002). Essentialism and folkbiology: Evidence from Brazil. *Journal of Cognition and Culture, 2*, 195–223.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science* (in press).

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 31*, 159–171.

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In W. G. Austin, & S. Worschel (Eds.). *The social psychology of intergroup relations* (pp. 33–47). Pacific Grove, CA: Brooks/Cole Publishing.

Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social Justice Research, 21*, 263–296.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited. *Psychological Bulletin, 116*, 21–27.

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics, 1*, 1–7.

Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. *Environmental impact assessment, technology assessment, and risk analysis* (pp. 107–129). Berlin Heidelberg: Springer.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the Fusiform Face Area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience, 23*, 3343–3354.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2016). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review,* 1–12.

Vorauer, J. D., & Sasaki, S. J. (2009). Helpful only in the abstract? Ironic effects of empathy in intergroup interaction. *Psychological Science, 20*, 191–197.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., ... Morey, R. D. (2016). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review* (in press).

Woolf, L. M., & Hulsizer, M. R. (2004). Hate groups for dummies: How to build a successful hate group. *Humanity and Society, 28*, 40–62.

Xu, F., & Rhemtulla, M. (2005). In defense of psychological essentialism. *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 2377–2380). NJ: Erlbaum Mahwah.

Zaki, J., & Cikara, M. (2015). Addressing empathic failures. *Current Directions in Psychological Science, 24*, 471–476.