UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Ekstrom, P., & Lai, C. (2022). A Good Person Shouldn't Feel This Way: Moralized Attitudes, Identity, and Self-Esteem. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.36344

Social Psychology

A Good Person Shouldn't Feel This Way: Moralized Attitudes, Identity, and Self-Esteem

Pierce Ekstrom[1] [a], Calvin Lai[2]

[1] Political Science, University of Nebraska - Lincoln, NE, US, [2] Psychological and Brain Sciences, Washington University in St Louis, St. Louis, MO, US

Keywords: Morality, Attitudes, Self-Esteem, The Self

https://doi.org/10.1525/collabra.36344

Collabra: Psychology

Vol. 8, Issue 1, 2022

Moralized attitudes are the attitudes that people construe as matters of right and wrong. In this study, we examine how moralized attitudes relate to how people evaluate themselves using the Attitudes, Identities, and Individual Differences (AIID) dataset—a survey of over 200,000 individuals asked to report their attitudes in one of 95 domains. In pre-registered analyses that were based on exploratory analyses of a subset of the data, we found that the specific attitudes that people moralize differ greatly from individual to individual and that moralized attitudes are more central to one's identity than non-moralized attitudes. We also examined whether mental conflict between identity-central attitudes and gut feelings about the corresponding attitude objects would be related to lower self-esteem, finding mixed and weak evidence supporting that claim. Together, our findings indicate that the attitudes that people moralize are tremendously diverse and are reliably connected to a sense of self. At the same time, peoples' self-esteem may be resilient to specific instances in which their gut feelings fall short of the attitudes that are central to their identity.

The attitudes that people moralize structure how they perceive and evaluate others (Skitka, 2010). Although people are willing to put up with disagreement about things that they see as matters of personal preference (e.g., attitudes toward broccoli), they are less tolerant of those who disagree with them about things that they see as matters of right and wrong (e.g., attitudes toward abortion).

In the current paper, we test the novel claim that moralized attitudes also structure how people perceive and evaluate themselves. We ground this claim in three existing areas of research. First, research on social perception has shown that perceptions of moral traits and characteristics dominate perceivers' evaluations of targets. (Fiske et al., 2007; Goodwin et al., 2014). Second, evidence dating back at least as far as Aronson's revised interpretation of Cognitive Dissonance Theory (1969) indicates that people are motivated to view themselves as moral and that morality is an important part of the self. Finally, Skitka and colleagues' (2010; 2005) work on moral conviction reveals that what constitutes "morality" varies across perceivers and has a powerful impact on social perceptions above and beyond the impact of other types of attitude strength. Integrating these lines of work, we propose that individuals' moralized attitudes comprise an important component of the self-concept that informs self-evaluation. Moralized attitudes may matter more for identity than non-moralized attitudes because they define whether people—including the self—are good or bad.

To test these possibilities, we first tested whether moralized attitudes are more central to individuals' identities than non-moral attitudes. Second, we tested whether identity-central attitudes may provide a rubric for self-evaluation. When participants' spontaneous feelings about a target contradicted what they believed their "real" attitude to be, we expected them to report lower self-esteem to the extent that their attitude was central to their identity. In sum, we tested the plausibility of a model in which moralized attitudes help define individuals' identities, and these identities in turn define the standards they use to evaluate themselves.

Morality, the Self, and the Motive to be Moral

People perceive and evaluate others based primarily on their moral traits and characteristics. In social perception, morality consistently outweighs competence (Fiske et al., 2007; Wojciszke, 1994, 2005; Wojciszke et al., 1998) and sometimes even interpersonal warmth (Landy et al., 2016). We expect morality to play a similarly dominant role in perception and evaluation of one's self. Consistent with this idea, a robust body of evidence indicates that people are motivated to perceive themselves as moral (Jordan et al., 2011; Klein & Epley, 2016; Merritt et al., 2010; Prentice et

a Please send correspondence to: pierce.ekstrom@gmail.com

al., 2019; Stanley et al., 2017; Tenbrunsel et al., 2010), suggesting that morality is a valued component of the self and an impactful component of identity (Aquino & Reed, 2002; McFerran et al., 2010; Reed & Aquino, 2003; Winterich et al., 2009; Wojciszke et al., 2011).

On the one hand, people exaggerate their own morality. They overestimate or over-report the morality of their past behavior and make similarly sunny and inaccurate predictions for their future behavior (Tenbrunsel et al., 2010). People also tend to claim that they are more moral than others—specifically, that they themselves are less likely to engage in immoral behaviors than others are (Klein & Epley, 2016). When people do acknowledge moral transgressions, they do not think those transgression were as bad as those committed by others (Stanley et al., 2017).

On the other hand, people feel threatened when their moral self-image is undermined. For example, the cognitive dissonance that people experience following counter-attitudinal advocacy has been partly attributed to the negative identity implications of having lied to a stranger for little external incentive (Aronson, 1969). People feel dissatisfied with recent life events to the extent that those events made them feel immoral, suggesting a close link between one's moral self-image and subjective well-being (Prentice et al., 2019). Threats to one's moral self-image can also drive moral action; studies have demonstrated that recalling past immoral behaviors can motivate people to behave more morally (Jordan et al., 2011).

In sum, although defensive biases in self-perception may often help individuals ignore or minimize their moral shortcomings, they are motivated to restore their moral self-images when these defenses are overcome. This motivated defense and pursuit of moral self-perceptions implies that morality plays a similar role in how people evaluate themselves as it does in how they evaluate others.

Individual Differences in Definitions of Morality

People differ in how they define "morality." Some cherish and protect tradition as a matter of principle. Others view many traditions as outdated or oppressive. Some value art and creative works as defining features of humanity. Others regard such activities as a waste of time and resources. Values of fundamental importance to some people's moral judgments can be irrelevant or even repugnant to others (Haidt, 2012; Janoff-Bulman et al., 2008; Schwartz & Bilsky, 1987). The moral yardstick by which people measure themselves and others therefore depends critically on their personal moral beliefs and values. These individual differences in what people choose to moralize have received considerable attention in research on political attitudes, where divergent moral beliefs and values abound. For example, some people view capital punishment as morally right and necessary, whereas others view it as morally wrong. Others do not view their support or opposition for the death penalty in moral terms. They may support the death penalty because they believe it deters crime or oppose it because expensive appeals processes drain public resources.

Work on moral conviction suggests a direct link between attitude moralization and perceptions of others. People who moralize their opinions on specific issues are uniquely intolerant of those who disagree with them about those issues (see Skitka, 2010 for review). They also express less willingness to work with or befriend moral dissenters (Skitka et al., 2005) and more willingness to openly discriminate against moral dissenters compared to those who disagree with them in non-moral domains (Wright et al., 2008). These effects of moral conviction persist even after controlling for other indices of attitude strength, such as importance, extremity, and certainty (Skitka, 2010).

This research demonstrates the value of an idiographic (i.e., participant-specific) approach to moral psychology—examining how individuals define morality for themselves reveals the powerful role that idiosyncratic values have in how they perceive others. Similarly, idiosyncratically moralized attitudes may influence individuals' perceptions of themselves. People are motivated to establish and maintain a moral self-image (Jordan et al., 2011), but the characteristics of a "good person" may differ from individual to individual.

The Current Study: Moralized Attitudes and Self-Evaluation

We examined whether the idiosyncratic attitudes that individuals moralize may inform their self-evaluations in the same way that broader, agreed-upon moral imperatives seem to color individuals' evaluations of both others and themselves.

Most people agree that murder is wrong and would think less of themselves if they had killed someone. Indeed, negative attitudes toward murder might be so strong and intensely moralized that even *feeling* inclined toward murder (in general or at a specific moment) might lead people to evaluate themselves more negatively. We hypothesized that any attitude imbued with moral significance could become a moral mandate with implications for self-evaluation. For example, people who dislike McDonald's because they prefer healthier foods may experience some momentary discomfort when they crave a Big Mac. Other people, who view their dislike of McDonald's as a reflection of their core moral values (e.g., vegetarians), might experience these cravings as a more serious threat to their identity. The inconsistency between their moralized attitude and spontaneous affective reactions could threaten their moral self-image and have consequences for their self-esteem.

These inconsistencies are not rare. Ex-smokers or ex-drinkers may crave cigarettes or alcohol despite years of abstinence. People deeply invested in their career success may feel an aversion to late afternoon or early morning work. A happily married person may feel sexually attracted to a stranger. And as any scholar of racism could attest, committed egalitarians can experience twinges of anxiety or aversion that connote racial prejudice (Gaertner & Dovidio, 1986). These spontaneous affective reactions may not escape individuals' notice, but they are outside of their immediate control. As a result, they can often conflict with the attitudes that individuals consciously espouse, even when those attitudes are of serious moral significance and central to individuals' identities (Gawronski & Bodenhausen, 2007; Ranganath et al., 2008).

We predicted that these attitude-inconsistent feelings should be troubling in moralized attitude domains, to the extent that moralized attitude domains are relatively central to individuals' self-definition. For example, a vegetarian who loves the taste of bacon may suffer more negative self-evaluations to the extent that they moralize and identify with anti-meat attitudes. In sum, we theorize that individuals' moralized attitudes will inform their self-evaluations because these attitudes define the kind of person an individual wants to be.

This theory is part of a larger body of research on which attitudes matter more than others—and *how* they matter. Attitude strength is the extent to which attitudes are stable over time, resist change, and impact behavior (Krosnick & Petty, 1995), but attitudes can meet these criteria in a number of ways. For example, attitude strength can reflect when people feel relatively certain about their attitudes, when they report their attitudes to be "important" to them, or when their attitudes are extreme (i.e., distant from neutrality). These multiple types of attitude strength (or "strength-related attitude attributes") are neither theoretically nor empirically reduceable to a single dimension of attitude strength (Visser et al., 2006). Attitude moralization—the extent to which someone perceives their attitude as a matter of right and wrong rather than personal preference—is one type of attitude strength. So too is attitude identity centrality—the extent to which someone perceives their attitude as part of who they are. Our theory concerns attitude moralization and identity centrality *per se,* distinct from other types of attitude strength.

We predicted that moralized attitudes would be uniquely central to individuals' identities, above and beyond how certain, important, or extreme they were. We also predicted that when spontaneous affective reactions and self-reported attitudes were inconsistent, people would report lower self-esteem to the extent that attitudes were central to their identity—again, above and beyond how certain, important, or extreme their attitudes were. Certain, important, and extreme attitudes may be central to individuals' identities, and individuals may feel uncomfortable when their spontaneous affective reactions contradict them, but our analyses distinguish moralization and identity centrality from these other types of attitude strength.

Overview and Hypotheses

We argue that individuals' beliefs about right and wrong define the identities that they wish to embody and the people that they want to be. We test these ideas using the Attitudes, Identities, and Individual Differences (AIID; Ebersole et al., 2019) dataset, an online study of over 200,000 individuals' reactions to attitude targets from 95 pairs that include social groups, political issues, individuals, objects, and abstract ideas (collected during 2004-2007; Summary of dataset available at: https://osf.io/pcjwf/). Because of the diverse attitude targets presented (each participant was randomly assigned to complete a survey on just one pair of 95 possible pairs), this dataset presents a unique opportunity to examine both aggregate-level characteristics of attitude domains and individual-level reactions to attitude targets.

We test two hypotheses with the current study:

1. **Moralized attitudes must be relatively central to identity (Morality-Identity hypothesis).** We predicted that moralized attitudes would comprise a central component of individuals' self-concepts. These attitudes would be part and parcel of how people see themselves—more so than non-moralized attitudes. We tested this hypothesis by predicting multiple measures of attitudes' identity centrality from the extent to which participants reported their attitudes to be connected to their personal values. When participants claimed their attitude was value-relevant, we expected them to identify relatively closely with that attitude and to identify with the target of that attitude. We predicted the effects of value relevance on identity centrality should hold above and beyond the effects of attitude importance, certainty, or extremity.
2. **Identities must provide a rubric for self-evaluation (Identity Rubric hypothesis).** We predicted that people would feel worse about themselves when their spontaneous affective reactions are inconsistent with identity-central attitudes. We tested this hypothesis by predicting participants' self-esteem from their gut feelings toward specific attitude objects, their attitudes' identity centrality, and the interaction between these variables. We focus on "gut feelings" about attitude stimuli because these relatively uncontrollable feelings may be a common type of attitude-inconsistent reaction. We include the interaction term because it allows us to quantify the extent to which participants report negative gut feelings specifically about identity-central attitude objects (or positive gut feelings about attitude objects they find inconsistent with their identity). We also expected these effects to hold above and beyond attitude importance, certainty, and extremity.

To test these hypotheses, we first conducted a series of exploratory analyses using a subset of the data (about 15% of the full dataset) that its curators had made publicly available for scholars to conduct preliminary analyses and form hypotheses. These exploratory analyses were reported in the Stage 1 submission for this registered report (https://doi.org/10.17605/osf.io/zry5b) and are included in the online supplemental materials for readers' reference (https://osf.io/6ckns/). They yielded strong support for the Morality-Identity hypothesis and weak and inconsistent support for the Identity Rubric hypothesis, but they lacked adequate power to test the latter.

We next conducted a pre-registered study to validate the measures employed in the AIID, some of which were designed for purposes different than our own.

We then repeated our exploratory analyses using the full AIID dataset, including both the exploratory subset of the data and the 85% of cases reserved for confirmatory tests. In addition to testing the Morality-Identity and Identity Rubric hypotheses, we conducted a descriptive analysis of which *types* of attitudes people tend to moralize or regard as central to their identities. Much existing work on the psychology of morality has taken a nomothetic (i.e., "objective," experimenter-driven) approach to morality, assuming

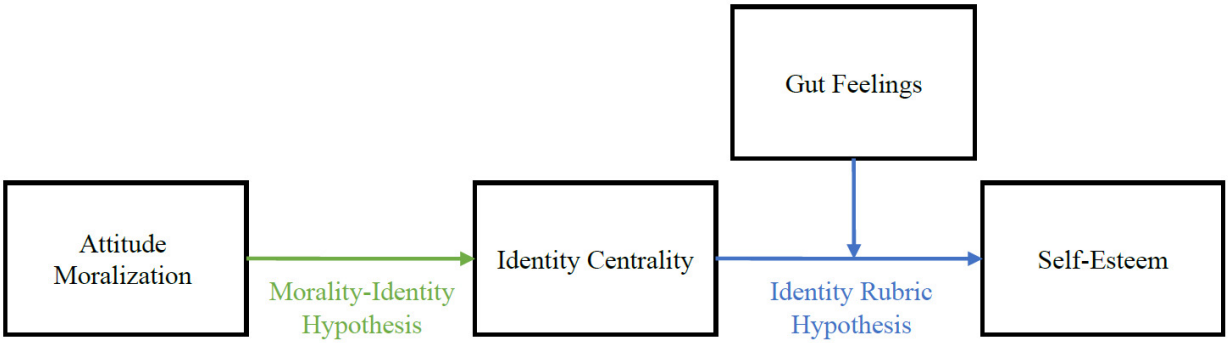

Figure 1. Illustration of hypotheses.

that certain patterns of thought or behavior (e.g., generosity) are consistently moralized across individuals (see Graham et al., 2012). If certain types of attitudes (e.g., toward abortion) are consistently moralized across individuals, then researchers can reliably use stimuli from these attitude domains to study the psychology of morality. If people vary widely in what they moralize (e.g., if many people do not actually moralize presumptively moral issues like abortion), then researchers may need to use a diverse set of attitudes to effectively model moral judgment. The more difficult it is to identify attitude domains that are consistently moralized across individuals, the more important it will be to employ an idiographic (i.e., subjective, participant-driven) rather than a nomothetic definition of morality in our analyses.

Because the AIID dataset is correlational, it cannot provide a definitive test of our larger causal theory (depicted in Figure 1). However, we *can* use the AIID data to subject the component paths of the theory to pre-registered tests. Evidence inconsistent with either the Morality-Identity hypothesis or the Identity Rubric hypothesis would require a fundamental revision to the theory.

In sum, the current paper reports (1) a novel study to validate measures employed in the AIID, (2) key descriptive statistics from the full AIID dataset, and (3) confirmatory results from the hypothesis tests we conducted using the full AIID dataset. Preliminary exploratory tests of the Morality-Identity hypothesis and the Identity Rubric hypothesis using the exploratory subset of the AIID data are presented in the online supplemental materials.

Transparent and Open Practices Statement

All data, analysis code, and research materials for this project are available on our project's OSF page: https://osf.io/6ckns/, including supplemental analyses mentioned but not reported in detail in the main text of this manuscript. The OSF page also includes power analyses we used to determine our sample size. Data were analyzed using Stata 17 (StataCorp, 2021). All analyses except those explicitly marked as exploratory were pre-registered at https://doi.org/10.17605/osf.io/zry5.

AIID Method

AIID Data

Our primary data came from the Attitudes, Identities, and Individual Differences (AIID) Study (see https://osf.io/pcjwf/ for documentation; Ebersole et al., 2019). Participants (total $N \approx 200{,}000$) were randomly assigned to answer questions about one of 95 attitude pairs (e.g., African Americans and European Americans, Burger King and McDonald's). These pairs were extremely diverse, offering a broad sample of the universe of targets toward which individuals might have attitudes, with some targets likely to be moralized by many people (e.g., Jews), some moralized by very few (e.g., Burger King), and others less uniformly viewed as moral or non-moral (e.g., United States). This variability makes the AIID a prime dataset for our idiographic approach to assessing attitude moralization and facilitates the mixed-effects "stimulus sampling" models that we employ (Judd et al., 2012).

The AIID used a planned missingness design, with some variables measured using different items across participants (e.g., moralization) and other variables sometimes not measured at all (e.g., self-esteem). We describe our strategies for dealing with missing data below. Table 1 summarizes the number of participants who completed each of our measures.[1]

AIID participants were a diverse set of online volunteers at the Project Implicit website. (63% women, 37% men; 72%

1 Note that these *N*s do not necessarily reflect the sample size in any given model or analysis. For clarity, we provide sample *N*s for each model reported in the results section.

Table 1. Number of Observed Cases per Measure

Variable	Exploratory Subset *N*		Confirmatory Subset *N*		Merged Dataset *N*	
Implicit Identification	7,640		38,405		46,045	
Self-Esteem	1,503		7,318		8,821	
	Target A	Target B	Target A	Target B	Target A	Target B
Gut feelings	31,472	31,482	152,318	152,362	183,790	183,844
Target Identity Centrality	13,550	13,554	65,005	64,992	78,555	78,546
Moralized Attitude Composite	31,168	31,174	150,660	150,627	181,828	181,801
Item 1-Negative judgments are wrong	7,709	7,705	37,863	37,870	45,572	45,575
Item 2-Positive judgments are wrong	7,843	7,843	37,665	37,657	45,508	45,500
Item 3-Negative judgments are acceptable	7,837	7,844	37,409	37,390	45,246	45,234
Item 4-Positive judgments are acceptable	7,779	7,782	37,723	37,710	45,502	45,492
Attitude Identity Centrality Composite	31,141	31,132	150,552	150,575	181,663	181,707
Item 1-Rejecting is inconsistent w/ self	7,862	7,865	37,663	37,687	45,525	45,552
Item 2-Accepting is inconsistent w/ self	7,814	7,814	37,356	37,360	45,170	45,174
Item 3-Rejecting is important to self	7,695	7,698	37,869	37,885	45,564	45,583
Item 4-Accepting is important to self	7,770	7,755	37,634	37,643	45,404	45,398

White, 6% Black, 6% Hispanic, 7% Asian, 1% Native American, 8% other or multi-racial; Age ranged from 7 to 88 years, M = 30, SD = 11.90). The sample was more educated and liberal than the U.S. population at large. Of those who reported their education (n = 269,951 total), 86% claimed at least some college education, and 56% of those who reported their political identity (n = 264,907) identified as liberal (compared to 26% for moderates and 18% for conservatives). Participants reported countries of residence across 6 continents, but the vast majority came from the United States (n = 214,445), Canada (n = 12,650), and the United Kingdom (n = 12,680). Data collection took place between 2004 and 2007.

The curators of the AIID dataset divided it into an exploratory dataset with 15% of AIID participants and a confirmatory dataset with the remaining 85%. Following our pre-registered analysis plan, the analyses we present below use a merged dataset that includes both subsets, but we reach the same conclusions when we rely on only the confirmatory subset of the data (see supplementary analyses: https://osf.io/6rq9h/).

AIID Procedure

Participants in the AIID first responded to a detailed battery of demographic questions to register themselves as Project Implicit users. Participants were then randomly assigned one of 95 pairs of attitude targets (see supplemental materials).

Next, in random order, participants completed an Implicit Association Test (IAT) and a self-report questionnaire focused on those attitude targets (Greenwald et al., 1998). The IAT was constructed *either* to assess implicit *evaluations*—associations between the attitude targets and "good" vs. "bad"—or to assess implicit *identification*—associations between the attitude targets and "self" vs. "other."[2] During the questionnaire, participants responded to a partly random[3] subset of 27-29 items from a pool of 76 possible items about the attitude targets, which assessed participants' explicit attitudes toward each target and several other attitude-related variables, such as attitude importance, certainty, and stability. Our key measures of moralized attitudes, identity centrality, and gut feelings were part of this questionnaire.

Finally, participants were randomly assigned to complete 1 of 20 individual difference measures. Rosenberg's (1965) Self-Esteem Scale was one of these possibilities.

AIID Measures

Attitude targets within pairs: "A" and "B." Each of the 95 attitude pairs included in the AIID data consisted of two attitude targets (e.g., African Americans and European Americans, Burger King and McDonald's). Targets included social groups (e.g., Muslims, Lawyers, Nerds), abstract ideas (e.g., Giving, Wisdom, Chaos), fixtures of popular culture (e.g., Julia Roberts, 50 Cent, Harry Potter), physical places (e.g., New York, Mountains), consumer products (e.g.,

2 In this article, we use "implicit" to refer to indirect measurement and use a theory-uncommitted conception of "association" (Greenwald et al., 2005; Greenwald & Lai, 2020). We do not make claims about the exact nature of the constructs measured by Implicit Association Tests.

3 See https://osf.io/pcjwf/ for randomization details.

Pepsi), political groups and policies (e.g., Democrats, Gun Rights), and health-related behaviors (e.g., Exercising, Drinking). The full list of attitude pairs and targets is included in our supplemental materials. Because of the wide range of attitude targets included in the dataset, we refer to the distinct targets within each pair as simply A and B. For each attitude pair, A and B were defined using average evaluative Implicit Association Test (IAT) scores in the AIID itself. Target A was always the one that participants implicitly preferred over the other on average.

In our tests of the Morality-Identity hypothesis, we can use A and B as repeated observations in multi-level models that examine the link between moralization and identity centrality, taking into account both within- and between-subject variance in these variables.

This multi-level strategy is not possible in our tests of the Identity Rubric hypotheses; self-esteem is the key outcome in those tests and is observed only once per subject. In this case, we simply ran two sets of models – one predicting self-esteem from gut feelings and identity-central attitudes toward A, the other predicting self-esteem from gut feelings and identity-central attitudes toward B.[4] We expect these models to yield substantively identical results.

Gut feelings and actual feelings. The AIID measured participants' gut feelings toward A and B after the following introduction: "People's gut feelings about a topic can be different from their feelings after they have had time to think about it. For example, someone who is trying to quit smoking might have a very positive gut feeling, but negative actual feelings toward smoking." Participants were then told to rate their *gut feelings* on a 10-point scale, from strongly negative (1) to strongly positive (10). Participants also reported their *actual feelings* on the same 10-point scale. Previous research indicates that reports of gut feelings and indirect, reaction-time measures (e.g., IAT scores) both measure something distinct from participants' controlled attitudes toward those targets (Ranganath et al., 2008). Whereas measures like the IAT rely on the speed of categorization to infer participants' spontaneous affective reactions, self-report measures of "gut feelings" rely on participants' *perceptions* of their spontaneous affective reactions. Both are useful indices of individuals' spontaneous affective reactions that may contradict or complement participants' general self-reported attitudes (see Gawronski & Bodenhausen, 2007). That said, our measures of gut feelings are better suited for our planned analyses than the IAT. Whereas the IAT can only measure spontaneous affective reactions relatively (i.e., target A vs. target B), our measures of gut feelings allow us to examine feelings toward targets A and B as separate predictors of self-esteem.

Moralized attitudes. All participants indicated the extent to which their attitudes toward A and B were tied to their personal values, providing an index of moralized attitudes. Each participant responded to only one item for each of the two attitude targets from among the following possible items, modified from Plant and Devine's (1998) internal motivation to respond without prejudice scale (IMS):

1. Because of my personal values, I believe that making negative judgments about A/B is wrong.
2. Because of my personal values, I believe that making positive judgments about A/B is wrong. (reverse-scored)
3. Because of my personal values, I believe that making negative judgments about A/B is acceptable. (reverse-scored)
4. Because of my personal values, I believe that making positive judgments about A/B is acceptable.

Response options ranged from 1 (Strongly disagree) to 6 (Strongly agree). The same item (1, 2, 3, or 4) was always used to assess the moralization of A and B.

None of these items are a pure measure of attitude moralization *per se*. An ideal measure of attitude moralization would separately assess participants' attitudes and the extent to which they see those attitudes as a matter of right and wrong. Instead, the AIID asked participants to what extent it is right (or wrong) to hold a specific attitude which may or may not reflect their own. Thus, participants' responses to the AIID attitude moralization questions reflect both the direction of their personal attitudes and the moral significance of those attitudes (i.e., how "wrong" or "acceptable" they are "because of [their] personal values.") In our validation study (described below), we assessed whether participants' scores on these items were systematically related to the extent that they moralized the measured attitude.

To maximize our number of analyzable cases, some analyses reported in the Stage 1 submission of our manuscript used a *moralized attitude composite*. Each participant's score on this composite was equal to their response to a single item—whichever moralized attitude item they were assigned (with responses to Items 2 and 3 reverse-scored). Because none of the four moralization items were ever administered together, this composite helps us "fill in" missing data, but with a serious risk: we cannot assess their internal consistency or the reliability of the composite with the AIID alone. We therefore assessed the reliability of this composite in our validation study.

Identity centrality. The AIID included three distinct measures of identity centrality (IDC). We include all three to assess the robustness of our effects across distinct operationalizations of IDC. First, participants indicated each *target's identity centrality*: "How much is A/B part of your self-concept?" This "target IDC" measure indicates the extent to which participants identify with the target *per se*.

4 A single model that aggregated variables across A and B was not feasible due to diversity across attitude pairs. For example, the identity centrality scores for some pairs of attitude targets (e.g., Meg Ryan and Julia Roberts) are highly related in a positive direction, reducing the variance available for difference scores. This high collinearity also precludes us from safely entering A- and B-related variables as predictors in a single model. Meanwhile, the identity centrality scores for other pairs of targets are virtually unrelated (e.g., Effort and Talent, Innocence and Wisdom), making mean ratings meaningless.

Second, participants indicated their *attitudes' identity centrality*. Each participant responded to only one item for each of the two attitude targets from among the following possible items, modified from Plant and Devine's (1998) IMS:

1. Being rejecting of A/B is inconsistent with my self-concept.
2. Being accepting of A/B is inconsistent with my self-concept. (reverse-scored)
3. Being rejecting of A/B is important to my self-concept. (reverse-scored)
4. Being accepting of A/B is important to my self-concept.[5]

Response options ranged from 1 (Strongly disagree) to 6 (Strongly agree). This "attitude IDC" measure indicates the extent to which participants identify with specific positive or negative attitudes to the target.

Here too, we created a composite variable. Each participant's composite attitude IDC score was equal to their response to a single item—whichever attitude IDC item they were assigned (with Items 2 and 3 reverse-scored). Again, we cannot compute Cronbach's α for the attitude IDC composite because each participant only answered one of the four items. We used the validation study to assess their reliability.

Note that each attitude identity centrality item corresponds most closely with the moralization item of the same number. We employ these corresponding pairs in some of the analyses presented below.

Finally, a random subset of participants (about 25% of the full sample) completed an identity IAT to assess *implicit identification*. Whereas the more common evaluation IAT assesses implicit associations between two target categories and the concepts "good" and "bad," an identity IAT assesses implicit associations between two target categories and the concepts of "self" and "other." Higher *D* scores indicate a stronger implicit association between the self-concept and A, compared to the self-concept and B.

Self-esteem. A randomly determined subset of participants (N = 8,821; about 5% of all cases) were assigned to complete the Rosenberg (1965) Self-Esteem Scale(M = 4.69, SD = 0.95, α = 0.89). Possible scores ranged from 1 (minimum self-esteem) to 6 (maximum self-esteem). This served as our measure of participants' self-esteem.

Other Strength-Related Attitude Attributes. The AIID included one item to assess attitude importance ("How personally important are your feelings toward [A/B]?" from 1 "Not at all important" to 6 "Very important") and one item to assess attitude certainty ("How certain are you about your feelings toward [A/B]?" from 1 "Not at all certain" to 6 "Very certain"). To quantify attitude extremity, we computed the distance between the scale midpoint and participants' "actual feelings" toward the attitude target.

Validation Study

Because the AIID items measuring moralized attitudes and attitude identity centrality confound attitude valence with moralization and identification (respectively), we collected new data to test how well the AIID items measure these core constructs independent of valence. These data also allow us to perform reliability analyses impossible with the AIID alone and to examine the correlates of other key variables (i.e., target identity centrality, implicit identification).

Validation Study Participants

Participants were volunteer visitors to the Project Implicit website (https://implicit.harvard.edu) and therefore drawn from a population of individuals as similar as possible to those who comprised the AIID sample. We aimed to recruit enough participants that at least 800 individuals would complete our study, and 882 did so.[6] See the supplemental materials for relevant power analysis.

Consistent with the original AIID sample, our validation sample included more women (59%) than men (39%), more liberals (54%) than conservatives (21%), and more White than Non-White respondents (67% White, 8% Black, 9% Hispanic, 10% Asian, <1% Native American, 5% other or multi-racial). Respondents were relatively educated (90% with at least some college). Their reported age ranged from 15 to 110 years (M = 41, MDN = 40, SD = 14.06).[7] Participants were mostly from the United States (n = 668), Canada (n = 33), and the United Kingdom (n = 43). Data were collected in August 2021.

Validation Study Procedure

Each participant was randomly assigned to answer questions about one pair of stimuli from the AIID. All participants first completed an identification IAT for that stimulus pair, then reported their explicit attitudes toward each stimulus. Next, participants completed each of the following measures in random order: AIID moralized attitudes, AIID attitude identity centrality, AIID target identity centrality, a standard measure of attitude moralization (Skitka & Morgan, 2014), a standard measure of attitude identity centrality (Luhtanen & Crocker, 1992), attitude importance (Boninger et al., 1995). Participants completed each mea-

5 Our Stage 1 registered report incorrectly reported that Items 3 and 4 asked participants whether rejecting or accepting A/B was "consistent with [their] self-concept." This was a transcription error on our part. These items asked participants whether rejecting or accepting A/B was "important to [their] self-concept." Unfortunately, the transcription error also affected our validation study, in which we used "consistent with" rather than "important to."

6 These descriptive statistics are based on the participants who reached the final (debriefing) page of the study.

7 Because we did not pre-register any plan to exclude participants reporting unlikely demographic characteristics, we retained all participants regardless of age.

sure for both attitude stimuli (A and B) before proceeding to the next measure.

Validation Study Measures & Materials

The full validation study questionnaire is provided in Appendix A.

Explicit attitudes. Participants reported "How positive or negative do you feel towards [A/B]?" on a scale from 1 to 10.

AIID moralized attitudes. Each participant completed all 4 "moralized attitude" items from the AIID for both attitude targets (e.g., Because of my personal values, I believe that making negative judgments about [A/B] is wrong).

AIID attitude identity centrality. Each participant completed all 4 "attitude identity centrality" items from the AIID for both attitude targets (e.g., Being rejecting of [A/B] is inconsistent with my self-concept).

Due to a transcription error on our part, two attitude IDC items differed from their original wording in the AIID questionnaire. In our validation study, Items 3 and 4 asked participants whether rejecting or accepting A/B was "consistent with [their] self-concept." The original AIID asked whether rejecting or accepting A/B was "important to [their] self-concept." Follow-up analyses suggest that this error probably did not have a large impact on our results.[8]

AIID target identity centrality. We measured attitude targets' identity centrality using the same item used in the AIID (i.e., "How much is [A/B] part of your self-concept?").

Standard measure of attitude moralization. To measure attitude moralization independent of attitude valence, we used Skitka and colleagues' (e.g., Skitka & Morgan, 2014) *moral conviction* scale, which includes items such as "To what extent are your feelings about [A/B] a reflection of your core moral beliefs and convictions?" Any attitude, positive or negative, can be relatively high or low on this measure.

Standard measure of attitude identity centrality. To measure attitude identity centrality independent of attitude valence, we used modified items from Luhtanen & Crocker's (1992) collective self-esteem scale, which includes items such as, "My feelings about [A/B] are an important reflection of who I am." Any attitude, positive or negative, can be relatively high or low on this measure.

Attitude importance. Participants indicated the importance of their attitudes toward A and B using two items used by Boninger et al. (1995).

Attitude target stimuli. Participants were randomly assigned to answer questions about one of the following stimulus pairs, which we chose to represent the diverse attitude targets used in the original dataset: Tall People – Short People; Jews – Christians; Strong – Sensitive; Career – Family; Tax Reductions – Social Programs; Security – Freedom; Briefs – Boxers; Urban – Rural; Night – Morning; Pepsi – Coke.[9]

Validation Study Results (Pre-Registered Analysis)

Our validation study sought to assess the internal consistency and validity of the AIID's moralized attitude items and attitude identity centrality items. This study also enabled us to examine the correlates of target identity centrality and implicit identification.

Evaluating the AIID's moralized attitude measures. First, we assessed the internal consistency of the AIID's moralized attitude measures, to determine whether we can reasonably combine these items in a single scale. Results suggest that we cannot. We computed a separate α across the four moralized attitude items for each target stimulus (e.g., Tall People, Short People, Jews, Christians, Strong). Our pre-registered decision rule was that if the median of the resulting 20 αs was below 0.6, we would drop all proposed analyses using the "composite" measure of moralized attitudes. The median of the observed αs was 0.54, so we have dropped all proposed analysis using the "composite" measure of moralized attitudes.

We next assessed the validity of the AIID's moralized attitude measures. We estimated multi-level models predicting responses to each of the four items from moral conviction, explicit attitudes, and the interaction between these two predictors, controlling for importance and its interaction with explicit attitudes. This model included two observations for each participant—one for Target A, the other for Target B—and many participants assigned to each stimulus pair (e.g., Tall People – Short People). Therefore, we allowed the model intercept and slopes to vary randomly

8 Participants in the validation study were about half a scale point more likely to report judgments to be "consistent with" their self-concept (for negative judgments, $M = 2.45$, $SD = 1.30$; for positive judgments, $M = 4.00$, $SD = 1.52$) than participants in the AIID study (responding to the same stimuli) were to report judgments to be "important to" their self-concept (for negative judgments, $M = 1.99$, $SD = 1.40$; for positive judgments, $M = 3.43$, $SD = 1.86$). However, responses to both versions of each item were similarly correlated with other indices of identity centrality. The extent to which participants saw negative judgments as "important to" or "consistent with" their self-concept were both weakly correlated with target identity centrality ($r_{important} = 0.05$; $r_{consistent} = -0.05$) and implicit identification with the target ($r_{important} = 0.06$; $r_{consistent} = -0.07$). The extent to which participants saw positive judgments as "important to" or "consistent with" their self-concept were both moderately correlated with target identity centrality ($r_{important} = 0.65$; $r_{consistent} = 0.43$) and weakly with implicit identification with the target ($r_{important} = 0.09$; $r_{consistent} = 0.10$).We can only speculate, but the modest differences between these correlations suggest that the incorrectly worded items we used in the validation study would not differ fundamentally from the correctly worded items in how they relate to the benchmark we used to assess their validity.

9 Specifically, we selected these targets by first splitting the AIID stimuli into 4 post-hoc categories (described in detail below): Groups, Politics, Abstract Ideas, and Everyday Targets. We then randomly selected 4 attitude pairs from each category (8 from the larger Everyday Targets category) as potential stimuli. From those potential stimuli, we selected the 2 pairs (4 pairs for Everyday Targets) that we believed to best represent each category, automatically ruling out outdated pairs related to pop culture or contemporary events (e.g., Meg Ryan – Julia Roberts; George W. Bush – John Kerry).

across participants and across stimulus pairs except when estimating random slopes prevented model convergence.[10]

If the AIID items actually tap "moralization," then they ought to be related to moral conviction. Specifically, we predicted that participants with strong moral conviction (compared to low moral conviction) would perceive attitudes *opposed* to their own as more "wrong" and less "acceptable." At the same time, they should perceive attitudes *consistent* with their own as more "acceptable" and less "wrong".

We tested this using the conditional effects (i.e., simple slopes) of moral conviction. The effect of moral conviction on the AIID "acceptable" and "wrong" ratings ought to depend on the participant's attitude and the item used to assess it. For example, among participants who like Christians, moral conviction should positively predict the belief that "negative judgments about Christians are wrong." Moral conviction should negatively predict the same belief among participants who dislike Christians.

However, our validity evidence was mixed. See Figure 2. Item 1 ("Because of my personal values, I believe that making negative judgments about [A/B] is wrong") behaved as expected, but effect sizes were small. Participants with positive attitudes toward the target (8 on the 10-point scale) were more likely to agree that negative attitudes were wrong to the extent that they saw their positive attitude as a moral conviction ($b_{\text{moral conviction}}$ = .15, p = 0.021). Participants with negative attitudes (3 on the 10-point scale) were more likely to *disagree* that negative attitudes were wrong to the extent that they saw their negative attitude as a moral conviction ($b_{\text{moral conviction}}$ = -.15, p = 0.049).[11]

Items 2, 3, and 4, behaved as expected only for certain attitudes. Moral conviction predicted whether participants saw their own attitude (positive or negative) as *acceptable*—evident in the dashed line in Panel 3 (b = 0.23, p = 0.001) and the solid line in Panel 4 (b = 0.15, p = 0.003). Moral conviction also predicted whether participants saw the opposite attitude (positive or negative) as *wrong*—evident in the solid line in Panel 1 (b = 0.15, p = 0.021) and the dashed line in Panel 2 (b = 0.31, p < 0.001). However, Items 2 and 3 were unrelated to moral conviction among participants with positive attitudes, who uniformly disagreed that their positive attitudes were "wrong" (solid line, Panel 2; b = -0.01, p = 0.892) and that negative attitudes were "acceptable" (solid line, Panel 3; b = 0.05, p = 0.358). And Item 4 was unrelated to moral conviction among participants with negative attitudes, who were uniformly neutral as to whether positive attitudes were acceptable (dashed line, Panel 4; b = -0.02, p = 0.732). In sum, Item 1 was related to moral conviction as expected, but weakly. Items 2 and 3 were related to moral conviction for participants with negative attitudes, and Item 4 was related to moral conviction for participants with positive attitudes.

This evidence is weaker than we expected. On the one hand, we could reasonably argue that the AIID items are not valid measures of moralized attitudes. They are weakly intercorrelated, inconsistently related to the "gold standard" measure of moral conviction we used, and even when they are related to moral conviction, those relations are weak. On the other hand, we could also make the opposite argument. The four items may be weakly intercorrelated because they measure four distinct beliefs: whether positive judgments are acceptable, whether positive judgments are wrong, whether negative judgments are acceptable, and whether negative judgments are wrong. Each of these beliefs, in turn, was at least *sometimes* related to moral conviction, with some items (1 and 4) related to moral convictions about positive attitudes and others (2 and 3) related to moral convictions about negative attitudes. Moreover, our pre-registered reliance on the moral conviction as the gold standard may be overly narrow. We have defined moralized attitudes as the attitudes that people construe as matters of right and wrong. Although the term "acceptable" is arguably ambiguous in whether it invokes social norms or moral principles, the belief that certain judgments are "wrong" is by definition, a moral belief, regardless of how strongly it is correlated with reported moral conviction.

So long as the latter argument *can* be made—that the AIID items are at least partly valid indices of moralization for certain attitudes—we see some merit in proceeding with our pre-registered analyses. However, given the evidence that some items are more valid than others for certain attitudes, we must pay particular attention to whether our results differ across Items 1, 2, 3, and 4 (with analyses using Item 1 being the most informative tests of our hypotheses). Fortunately, the results of our subsequent analyses do not appear to depend which item we use, suggesting that if we focused our interpretations more narrowly (on, say, analyses using only Item 1), we would reach the same conclusions.

Evaluating the AIID's attitude identity centrality measures. We used the same basic strategies to evaluate the AIID's identity centrality measures. To assess their internal consistency, we again computed separate Cronbach's αs for each of our 20 target stimuli. This time, the median of the resulting 20 αs was 0.71, leading us to retain proposed analyses using the "composite" measure.

To assess the validity of the AIID's attitude identity centrality measures, we estimated multi-level models predicting responses to each of the four items and their composite. Predictors included Luhtanen and Crocker's (1992) measure of identity centrality, explicit attitudes, and the interaction

10 In our pre-registered analysis plan, we neglected to say that the slopes would also be allowed to vary randomly. This was an error. Following a suggestion from an anonymous reviewer, we revised *all* multi-level models in our paper to include random slop estimates, but we overlooked the analysis plan for our validation study when making that revision.

11 We chose to estimate conditional effects at attitude scores of 3 and 8 because 3 represented a moderately negative attitude and 8 represented a moderately positive attitude. Effects of moral conviction on item endorsement were generally smaller for neutral attitudes and larger for more extreme attitudes.

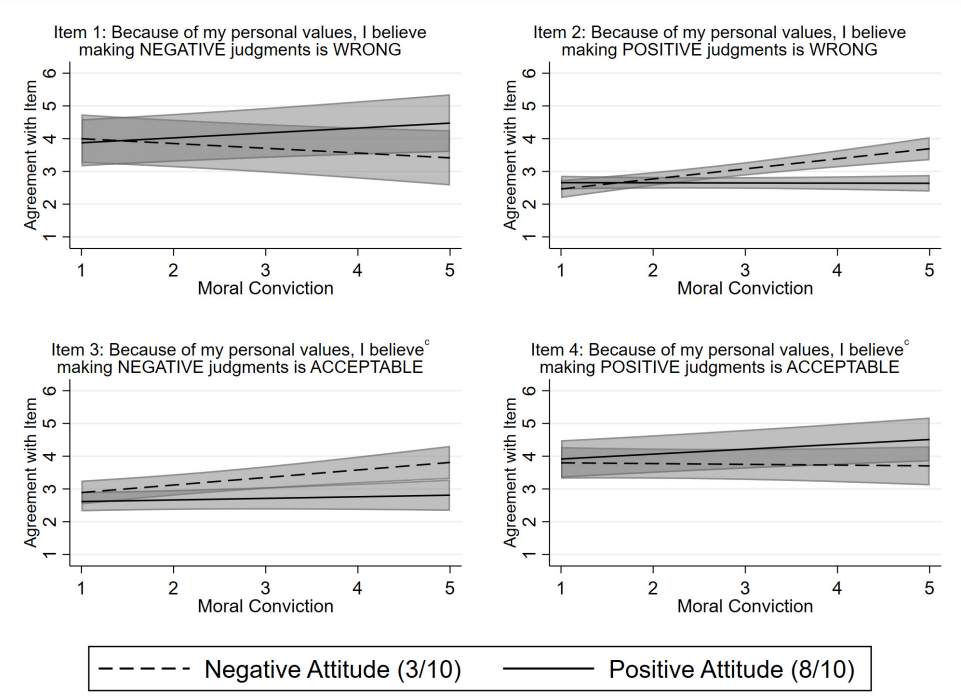

Figure 2. Lines indicate the predicted agreement with the relevant item across the full range of participants' moral convictions. Error bands indicate 95% CIs. Predictions derive from the fixed portion of four separate linear mixed models. All models included random slopes that varied across attitude stimuli and random intercepts that varied across stimuli and across participants. [c] Models predicting Items 3 and 4 did not converge when slopes were allowed to vary across participants, so these models (and only these models) did not include random slopes across participants.

between these two predictors. We did not include other covariates. Once again, because this model included two observations for each participant and many participants per stimulus pair, we allowed the model intercept and slopes to vary randomly across participants and across stimulus pairs.

Again, our assessment of each item's validity hinged on conditional effects. We predicted that people who scored relatively high on the Luhtanen and Crocker items (i.e., who describe their attitudes as an important part of who they are) should perceive attitudes opposed to their own as especially "inconsistent" and less "consistent" with their self-concept. They should show the opposite pattern for attitudes that align with their own.

And once again, evidence of the AIID items' validity varied depending on the item and on participants' attitudes. See Figure 3. Items 2 and 3 show the crossover interaction we expected and therefore provide the most valid indices of attitude identity centrality, although these items were more useful for assessing the identity centrality of negative attitudes (assuming a negative attitude of 3 out of 10, Item 2: $b_{\text{identity centrality}} = 0.13$, $p = 0.006$; Item 3: $b_{\text{identity centrality}} = 0.19$, $p < 0.001$) than of positive attitudes (assuming a positive attitude of 8 out of 10, Item 2: $b_{\text{identity centrality}} = -0.05$, $p = 0.064$; Item 3: $b_{\text{identity centrality}} = -0.03$, $p = 0.356$). Items 1 and 4, meanwhile, appear only to measure the identity centrality of positive attitudes.

In sum, AIID items 2 and 3 can assess the identity centrality of both positive and negative attitudes, but especially negative attitudes, whereas Items 1 and 4 are useful for assessing the identity centrality of negative attitudes but not positive attitudes. Once again, the results of our subsequent analyses were entirely consistent across items, suggesting that differences across items do not affect our conclusions.

Moreover, the alphas for these attitude identity centrality items suggest that the items can be safely combined in a composite scale. The validity study suggests that we can safely and productively use the composite measure of attitude identity centrality.

Target identity centrality and implicit identification. Our validation study also included the AIID's measures of target identity centrality and implicit identification. In our pre-registration, we wrote that these variables' "correlations with the AIID attitude identity centrality items will provide secondary evidence of the latter's validity." Because the extent to which someone identifies with some target (e.g., identifies as a tall person) is distinct from their identification with their attitudes toward a target (e.g., identifying as someone who likes tall people), and because the IAT is a comparative measure (e.g., of associating the self more with tall people than with short people), high correlations here are neither necessary nor sufficient to indicate the validity of the AIID measures of attitude identity centrality. Still, they may provide useful context for how much (or how little) these constructs have in common.

Target identity centrality was moderately associated with the forward-coded attitude identity centrality items ("Being accepting (rejecting) of [A/B] is consistent (incon-

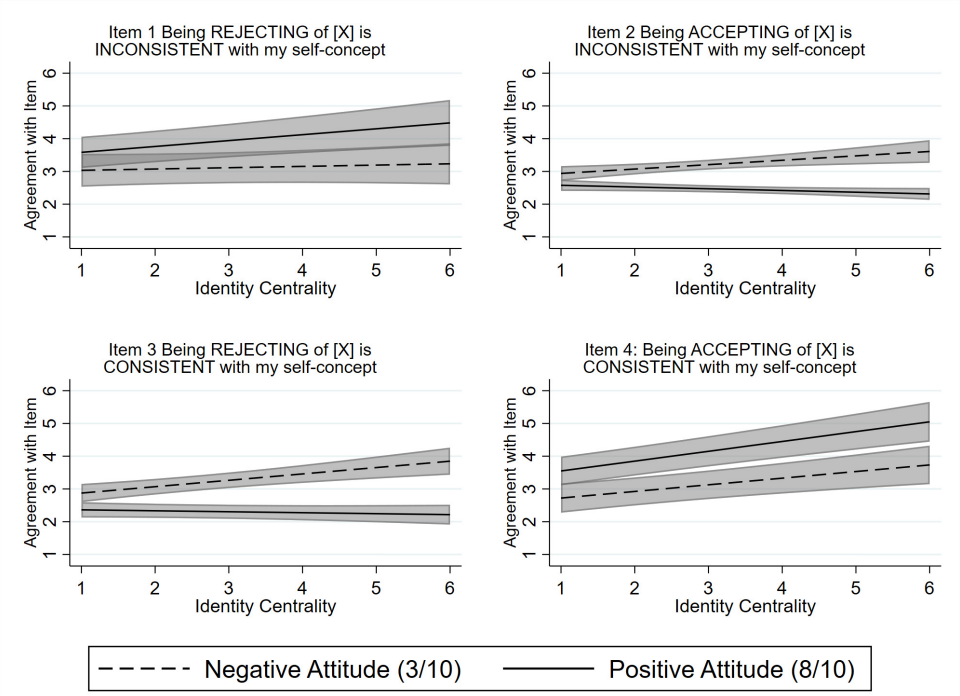

Figure 3. Lines indicate the predicted agreement with the relevant item across the full range of attitude identity centrality, measured using the modified Luhtanen & Crocker (1992) scale. Error bands indicate 95% CIs. Predictions derive from the fixed portion of four separate linear mixed models. All models included random slopes and intercepts that varied across stimuli and across participants

sistent) with my self-concept"; *r*s = .18, .32, .40, .46; all *p*s < 0.001) but weakly and inconsistently related to the reverse-coded items ("Being accepting (rejecting) of [A/B] is inconsistent (consistent) with my self-concept"; *r*s = -.14, -.07, .00, .00, *p*s = <.001, .034, .92, .92.). This finding underscores the fact that identifying with something is distinct from identifying as someone who likes that thing, even though the two may sometimes be related. The two variables are therefore not interchangeable in our analysis.

Identification IAT scores were not consistently correlated with target identity centrality, the identification IAT's most direct explicit analogue in our study (for Target A, $r = 0.25$, $p < 0.001$;for Target B, $r = 0.02$, $p = 0.39$). In hindsight, this is probably because very few of the AIID stimuli are polar opposites; people can simultaneously identify with Pepsi and Coke, with security and freedom, with "strong" and "sensitive" behaviors, with career and family life, etc. Perhaps for similar reasons, IAT scores were also quite weakly associated with attitude identity centrality items (-.11 < *r*s < .15). In short, attitude identity centrality, target identity centrality, and implicit identification with Target A over Target B are best understood as 3 distinct constructs in our analyses.

Results from the AIID Dataset (Pre-Registered Analysis)

Here we present the results of our pre-registered analyses of the AIID dataset. All files necessary to reproduce our analyses are provided in our online supplemental materials, accessible at the following link: <https://osf.io/6ckns/>. These files include the raw and processed data files, the scripts we used to conduct the analysis (in Stata 16 and 17), and a codebook that indicates which variables in the dataset correspond to those in the manuscript.

All of the results we report below are pre-registered reanalyses, except those from the models summarized in Table 9, which were pre-registered but never estimated using the exploratory data. We initially estimated each of these models using 15% of the AIID data. These exploratory analyses were conducted prior to registration and in-principle acceptance and were reported in our Stage 1 submission (https://doi.org/10.17605/osf.io/zry5b). The analyses we present here—following our pre-registered plan—use the full AIID dataset, including the 15% of the data we used initially and the 85% that we reserved for confirmatory tests alone. Exploratory and confirmatory results were almost uniformly consistent.

Although our pre-registration indicated that we would combine the 15% and 85% subsets of the AIID data for a single analysis, this decision blurs the boundary between data used for exploratory and confirmatory purposes. We therefore re-estimated all models used to test our Morality-Identity and Identity Rubric hypotheses using the 85% confirmatory dataset only. The size and statistical significance of coefficients were virtually unchanged, with two exceptions described below. They are described in detail in supplemental analyses available here: https://osf.io/6rq9h/.

Descriptive Analyses: Moralization and Identity Centrality Across Domains and Targets

We begin with a descriptive analysis. Were certain attitude targets more frequently moralized or identity-central than others? If so, future work might rely on these targets (or targets like them) to reliably invoke moralized attitudes. If not, then future work would benefit from allowing participants to articulate their own moral priorities. We made no firm predictions here.

Given the large number of attitude targets included in the AIID study (187 unique attitude targets across 95 attitude domains), we group them into 4 ad-hoc categories: Groups, Politics, Abstract Ideas, and Everyday Targets. Targets that referred to humans in the plural or as a collective were categorized as groups, unless those groups were defined by political ideology or partisanship. Targets that referred to policies, politicians, groups, or larger ideas associated with the political left or right were set in the Politics category. Targets that referred to anything that participants could not directly observe and that did not qualify for the political category were categorized as abstract ideas. The remaining targets were placed in the "Everyday Targets" category, which consisted of individual people, places, and things, including consumer products, physical places, health-related stimuli, and pop culture phenomena. We anticipated that if any attitudes were generally low in moralization, it would be those in the Everyday Targets category.

Figure 4 depicts each attitude target's mean values across our three key variables: target identity centrality, attitude identity centrality, and moralized attitudes. Attitude targets in the figure are sorted from lowest to highest moralization of positive reactions. The figure reveals three important patterns. First, the general upward slope of the points in every panel indicates that when people moralized positive reactions to attitude targets, those positive reactions and the targets themselves also tended to be identity central (for attitude IDC, b = 1.08, p < 0.001, 95% CI [1.02, 1.14]; for target IDC, b = 1.11, p < 0.001, 95% CI [0.83 , 1.39]).[12] Thus, these mean scores provide macro-level evidence consistent with our Morality-Identity hypothesis. Second, there are clear differences in each variable across attitude domains, with positive reactions to social groups (M = 4.65, SD = 1.60) being more consistently moralized than positive reactions to abstract ideas (M = 4.36, SD = 1.57), everyday targets (M = 4.17, SD = 1.77), and political targets (M = 3.99, SD = 1.73).

A third important pattern is the relatively limited range of moralized attitudes. Positive (or non-negative) reactions to "Giving"—the most highly moralized abstract idea and probably the most overtly prosocial stimulus in the AIID dataset—had an average moralization score of 4.93 (SD = 1.36) out of 6. That is, the average participant "agreed" with the assertions that positive reactions to giving are acceptable and that negative reactions to giving are wrong and "disagreed" with the inverse of these assertions (i.e., that negative reactions to giving are acceptable, that positive reactions to giving are wrong). Positive (or non-negative) reactions to McDonald's—the least moralized attitude domain in the Everyday Targets category—had an average moralization score of 3.12 (SD = 1.82). This score falls between "slight agreement" and "slight disagreement". If we round down, the average participant tended not to moralize positive reactions to McDonald's. The difference between these highest and lowest aggregate moralization scores is significant ($t(637)$ = 13.78, p < 0.001) and yields a large standardized effect size (d = 1.13). However, the range of moralized reactions was surprisingly small in substantive terms, spanning less than two points on a six-point Likert scale.

Together, these results indicate that some attitude domains are reliably more moralized than others. Positive reactions to social groups were most consistently moralized, followed by reactions to abstract ideas, political targets, and everyday targets. However, the absolute range of average moralization was relatively small, spanning just one third of the scale's potential range. Thus, although some attitude domains are more likely to be moralized than others, the idiographic approach that we employ affords us an important advantage, expanding the observed range of attitude moralization and telling us which attitudes that individual participants see in moral terms. The resulting subjective, participant-driven definition of morality helps ensure that we will not dismiss as trash any participant's idiosyncratic moral treasure.

Testing the Morality-Identity Hypothesis

We next tested our Morality-Identity hypothesis. When participants saw their attitude as a reflection of their personal values, did they also see that attitude as part of their identity? Our measures of "identity" included A) targets' identity centrality (e.g., identifying with Christians), B) attitudes' identity centrality (e.g., identifying as liking Christians), and C) implicit identification (e.g., implicitly identifying with Christians compared to Jews).

To answer this question, we estimated a series of mixed linear models. Each model includes three levels with the two attitude targets (A and B) as the level-1 units, participants as the level-2 units, and randomly-assigned attitude pairs as the level-3 units. Each model allowed the effect of moralized attitudes and the intercept to vary randomly across participants and across stimulus pairs. Our central focus in every model is on the coefficient for moralized attitudes, which we expected to positively predict each outcome. Table 2 summarizes results.

Target identity centrality: Participants moralized attitudes toward identity-central targets. Consistent with the Morality-Identity Hypothesis, the moralized attitude items significantly predicted target identity centrality (bs = .18, .10, .12, .31, ps < .001, $N_{\text{participants}}$ between 19,517 and 19,672). See Table 3. Figure 5 illustrates these effects and makes clear that they were small in magnitude. For example, even participants who "strongly agreed" that positive

12 Regression coefficients derive from OLS regressions using a target-level dataset (i.e., one row for each unique attitude target).

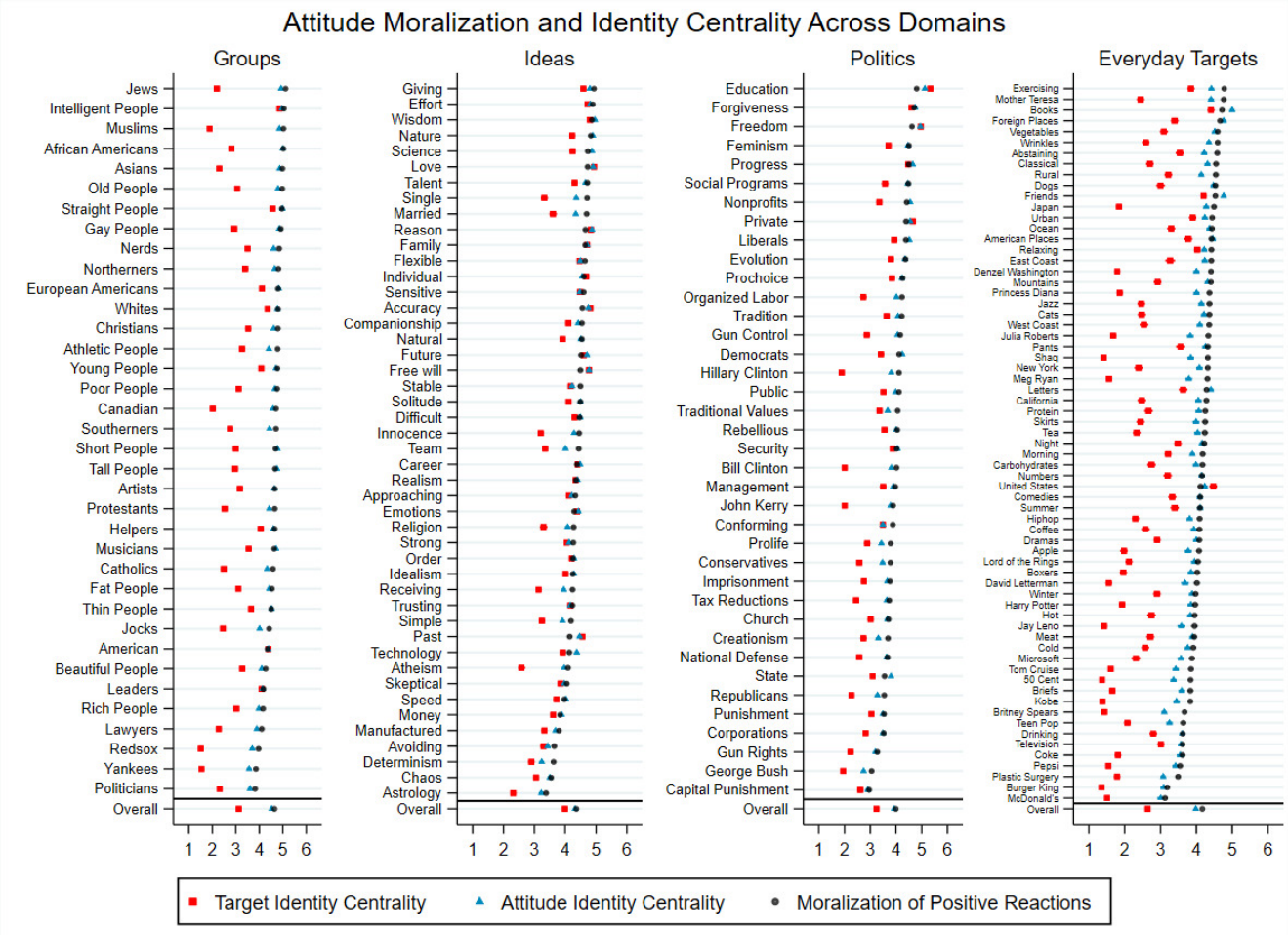

Figure 4. Figure depicts means and 95% CIs for our 3 key variables across all attitude targets. Moralization and Target IDC are composite items.

Table 2. Summary of Key Results

Testing the Morality-Identity Hypothesis: Did participants identify with the attitudes they moralized?		
Target Identity Centrality (Target IDC)	Attitude Identity Centrality (Attitude IDC)	Implicit Identification with Target A over Target B
Yes	Yes	Yes
These results held regardless of… which moralized attitude item we used to predict the identity outcome variable		
whether we controlled for attitude importance, certainty, extremity, or none of these.		
whether we used the full AIID dataset or the 85% reserved for confirmatory analysis.		

Testing the Identity Rubric Hypothesis: Did participants report lower self-esteem when their gut feelings were inconsistent with their identities?					
Target Identity Centrality (Target IDC)		Attitude Identity Centrality (Attitude IDC)		Implicit Identification with Target A over Target B	
For Target A?	Yes, slightly	For Target A?	Yes, slightly*	For Target A?	No
For Target B?	Yes, slightly*	For Target B?	No	For Target B?	No

*These results were significant in analyses of the full AIID dataset but not the 85% reserved for confirmatory analysis

judgments of A/B were acceptable or that negative judgments were wrong were relatively neutral as to whether A/B was a part of their self-concept. Still, these relations also held in separate models that controlled for attitude importance (*b*s for moralization = 0.10, 0.07, 0.09, 0.20, *p*s < .001, $N_{\text{participants}}$ between 6,417 and 6,599), certainty (*b*s for mor-

Table 3. Predicting Target Identity Centrality from Moralized Attitude Items (See Figure 5)

Outcome: "How much is [A/B] part of your self-concept?"	Item 1: Neg. judgments are wrong		Item 2 (R). Pos. judgments are wrong		Item 3 (R). Neg. judgments are acceptable		Item 4. Pos. judgments are acceptable	
Fixed Effects	*b* (S.E.)	*p*	*b* (S.E.)	*p*	*b* (S.E.)	*p*	*b* (S.E.)	*p*
Moralized Attitudes	0.18***	<0.001	0.10***	<0.001	0.12***	<0.001	0.31***	<0.001
	(0.01)		(0.02)		(0.01)		(0.02)	
Constant	2.53***	<0.001	2.71***	<0.001	2.71***	<0.001	1.76***	<0.001
	(0.08)		(0.08)		(0.08)		(0.08)	
Random Effects	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across participants)	0.09	(0.02)	0.07	(0.03)	0.09	(0.02)	0.08	(0.02)
σ^2 (Intercept across att. pairs)	0.64	(0.10)	0.59	(0.10)	0.58	(0.09)	0.50	(0.08)
σ^2 (Slope across participants)	0.01	(0.001)	0.01	(0.001)	0.01	(0.001)	0.01	(0.001)
σ^2 (Slope across att. pairs)	0.01	(0.002)	0.02	(0.003)	0.01	(0.002)	0.03	(0.004)
Log-likelihood	-70970		-71448		-71245		-69428	
Wald χ^2 (degrees freedom)	253.43*** (1)		37.39*** (1)		92.07*** (1)		358.08 *** (1)	
N Level-1 units (responses)	39,145		39,135		39,085		38,835	
N Level-2 units (participants)	19,672		19,654		19,639		19,517	
N Level-3 units (att. pairs)	95		95		95		95	

Note. Entries drive from 4 multilevel linear mixed models, each of which predicted Target Identity Centrality from a single moralized attitude variable, with random intercepts estimated across participants and across attitude pairs.
(†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

alization = 0.18, 0.10, 0.12, 0.30, *p*s < .001, $N_{participants}$ between 6,476 and 6,591), and extremity (*b*s for moralization =0.16, 0.08, 0.10, 0.27, *p*s < .001, $N_{participants}$ between 19,490 and 19,642).[13]

Attitude identity centrality: Participants moralized identity-central attitudes. We next estimated mixed models in which we regressed each attitude identity centrality item on the moralized attitude item with which it corresponded most closely (See Table 4 caption). Results were consistent with the Morality-Identity hypothesis (*b*s = 0.35, 0.45, 0.32, 0.48, *p*s < 0.001, $N_{participants}$ between 11,153 and 11,479). Moreover, the effects were large enough to be substantively meaningful. For example, participants who "strongly agreed" that positive judgments of A/B were acceptable or that negative judgments were wrong tended to agree (between "slightly agree" and "agree") that positive judgments were consistent with their self-concept and that negative judgments were inconsistent, and participants who disagreed with the former statements tended to disagree with the latter as well. See Table 4 and Figure 6. This pattern also holds in separate models that control for attitude importance (*b*s for moralization = 0.33, 0.43, 0.32, 0.39, *p*s ⩽ .001, $N_{participants}$ between 4,727 and 4,969), certainty (*b*s for moralization = 0.34, 0.45, 0.32, 0.50, *p*s < .001, $N_{participants}$ between 4,806 and 4,871), and extremity (*b*s for moralization = 0.34, 0.43, 0.31, 0.45, *p*s < .001, $N_{participants}$ between 11,134 and 11,461). This evidence suggests that in our sample, moralized attitudes tended to be more identity central than non-moralized attitudes, regardless of whether they were positive, negative, important, certain, or extreme.

Implicit identification: Participants moralized attitudes toward targets with which they implicitly identified. Finally, we examined participants' implicit identification with their attitudes. Because our implicit measure (an IAT *D* score) assessed identification with A *over* B, we created a difference score of the composite for moralized attitudes toward A minus the composite for moralized attitudes toward B. This difference score approach is not ideal; for many attitude targets, moralizing A does not imply viewing B as non-moral (e.g., wisdom vs. innocence). That said, the difference score allows us to analyze the robustness of our results across an additional measure of identity centrality

13 Controlling for all three of these variables simultaneously would reduce our sample size by about 93% , so we instead controlled for each in turn.

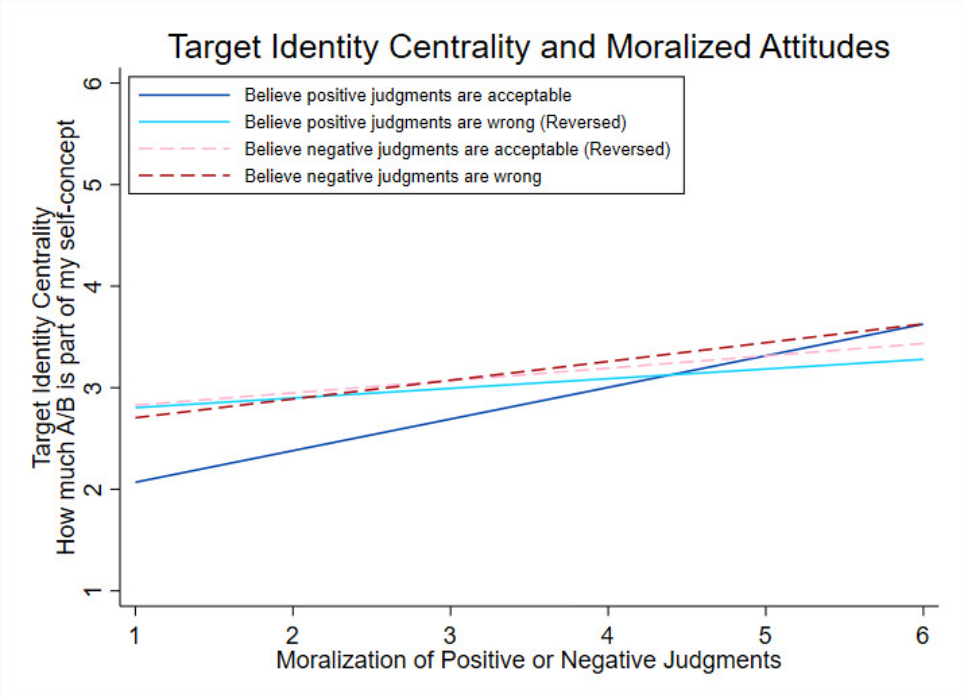

Figure 5. Lines indicate the predicted values of target identity centrality across the full range of each of the five moralization variables. These predictions derive from the fixed portion of four separate linear mixed models, each of which predicted IDC from a single moralized attitude variable with random slopes and intercepts that varied across participants and across attitude pairs. See Table 3.

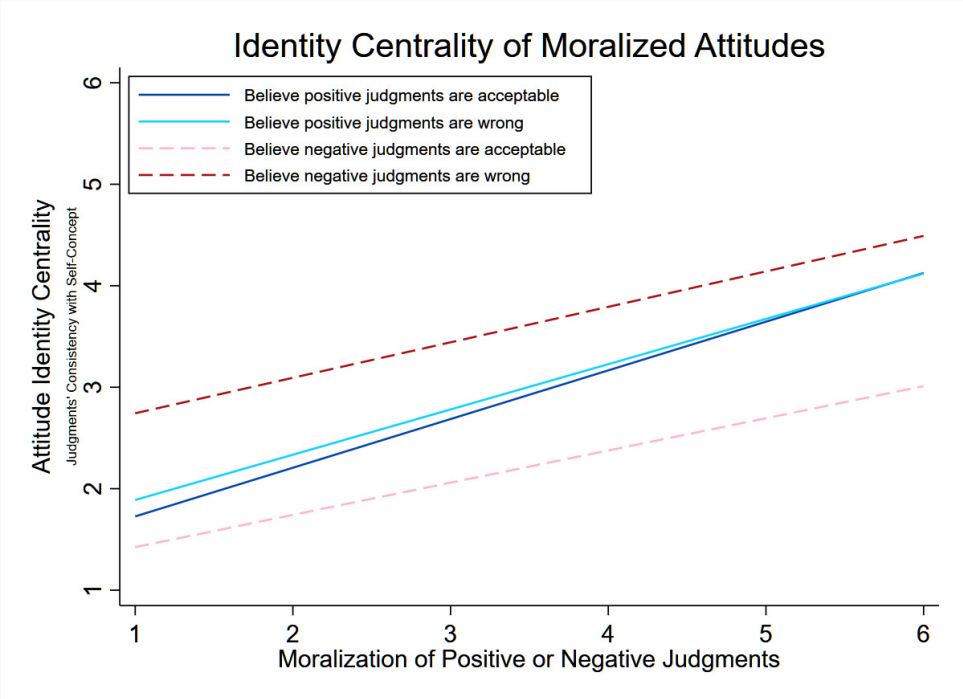

Figure 6. Lines indicate the predicted values of attitude identity centrality across the full range each moralization predictor. Predictions derive from the fixed portion of four separate linear mixed models, each of which predicted IDC from a single moralized attitude variable, with random slopes and intercepts that varied across participants and across attitude pairs. Table 4 includes details on which predictors were used for each attitude IDC outcome variable.

Table 4. Predicting Attitude Identity Centrality from Moralized Attitude Items (See Figure 6)

Outcome: Being [accepting/ rejecting] of [A/B] is [important to/ inconsistent] with my self-concept.	Item 1: Neg. judgments are wrong		Item 2 (R). Pos. judgments are wrong		Item 3 (R). Neg. judgments are acceptable		Item 4. Pos. judgments are acceptable	
	b (S.E.)	*p*	*b* (S.E.)	*p*	*b* (S.E.)	*p*	*b* (S.E.)	*p*
Fixed Effects								
Moralized Attitudes	0.35***	<0.001	0.45***	<0.001	0.32***	<0.001	0.48***	<0.001
	(0.01)		(0.01)		(0.01)		(0.02)	
Constant	2.39***	<0.001	1.44***	<0.001	1.11***	<0.001	1.25***	<0.001
	(0.05)		(0.03)		(0.02)		(0.08)	
Random Effects	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across participants)	0..98	(0.03)	0.60	(0.02)	0.26	(0.02)	0.70	(0.03)
σ^2 (Intercept across attitude pairs)	0.14	(0.03)	0.04	(0.01)	0.03	(0.01)	0.49	(0.09)
σ^2 (Slope across participants)	0.02	(0.002)	0.04	(0.003)	0.05	(0.002)	0.02	(0.005)
σ^2 (Slope across att. pairs)	<0.001	0.001	0.01	(0.003)	0.01	(0.002)	0.03	(0.005)
Log-likelihood	-42523		-38395		-36617		-39469	
Wald χ^2 (degrees freedom)	1375.16***(1)		911.58***(1)		768.84***(1)		669.58***(1)	
N Level-1 units (responses)	22,836		22,188		22,653		22,572	
N Level-2 units (participants)	11,479		11,153		11,377		11,343	
N Level-3 units (attitude pairs)	95		95		95		95	

Note. Entries are derived from multilevel linear mixed models. Outcomes differed across models, such that each moralized attitude item was regressed on the corresponding attitude identity centrality item. Corresponding pairs were as follows:

1. Because of my personal values, I believe that making negative judgments about A/B is wrong – Being rejecting of A/B is inconsistent with my self-concept (b = 0.36, p < 0.001, 95% CI [0.33, 0.37]).
2. Because of my personal values, I believe that making positive judgments about A/B is wrong – Being accepting of A/B is inconsistent with my self-concept (b = 0.45, p < 0.001, 95% CI [0.42, 0.48]).
3. Because of my personal values, I believe that making negative judgments about A/B is acceptable – Being rejecting of A/B is important to my self-concept (b = 0.32, p < 0.001, 95% CI [0.29, 0.34]).
4. Because of my personal values, I believe that making positive judgments about A/B is acceptable – Being accepting of A/B is important to my self-concept (b = 0.48, p < 0.001, 95% CI [0.44, 0.52]).

(†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

that does not rely on introspection and self-report. To the extent that participants did see any given target A as more or less morally charged than a given target B, we would predict that they would *also* associate A more closely with their self-concept.

Because each participant had only one score on these comparative measures, we estimated a *2*-level linear mixed model, with participants as the level-1 units and attitude domains as the level-2 grouping units. We examined the fixed effect of the moralized attitude difference score on implicit identification with A over B, allowing the slope and intercept to vary randomly across attitude pairs. We observed a small but significant relation between each moralized attitude difference score and implicit identification with A over B. The more strongly participants' attitudes were tied to their values, the more strongly those attitudes were tied to the self. See Table 5 and Figure 7. This pattern also held in separate models that control for attitude importance (*b*s for moralization = 0.04, 0.04, 0.05, 0.06, *p*s < .001, $N_{\text{participants}}$ between 4,402 and 4,522), certainty (*b*s for moralization = 0.05, 0.05, 0.04, 0.08, all *p*s < .001, $N_{\text{participants}}$ between 4,419 and 4,615), and extremity (*b*s for moralization = 0.04, 0.04, 0.05, 0.07, *p*s < .001, $N_{\text{participants}}$ between 10,289 and 10,617).

Testing the Identity Rubric Hypothesis

We next tested our Identity Rubric hypothesis: that self-esteem would be relatively low among participants whose gut feelings contradicted (A) their attitudes toward identity-central targets, (B) their identity-central attitudes, and

Table 5. Predicting Implicit Identification with A over B from Moralized Attitudes toward A vs. B (See Figure 7)

Outcome: Implicit identification with A over B	Item 1: Neg. judgments about A (vs. B) are more wrong		Item 2. Pos. judgments about A (vs. B) are less wrong		Item 3. Neg. judgments about A (vs. B) are less acceptable		Item 4. Pos. judgments about A (vs. B) are more acceptable	
	b (se)	*p*	*b* (se)	*p*	*b* (se)	*p*	*b* (se)	*p*
Fixed Effects								
Moralized Attitudes difference score	0.05*** (0.004)	<0.001	0.05*** (0.004)	<0.001	0.05*** (0.004)	<0.001	0.07*** (0.005)	<0.001
Constant	0.16*** (0.02)	<0.001	0.16*** (0.02)	<0.001	0.17*** (0.02)	<0.001	0.16*** (0.02)	<0.001
Random Effects	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across attitude pairs)	0.05	(0.01)	0.05	(0.01)	0.05	(0.01)	0.05	(0.01)
σ^2 (Slope across attitude pairs)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
Log-likelihood	-6656		-6671		-6459		-6623	
Wald χ^2 (degrees freedom)	110.97*** (1)		107.02*** (1)		180.58*** (1)		269.48*** (1)	
N Level-1 units (participants)	10,592		10,495		10,356		10,693	
N Level-2 units (attitude pairs)	95		95		95		95	

Note. Entries drive from 4 multilevel linear mixed models, each of which predicted Target Identity Centrality from a single moralized attitude difference score, with random intercepts and slopes estimated across attitude pairs. Every difference score was computed such that higher values indicate moralizing more positive (or less negative) reactions to A relative to B.
(†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

(C) their attitudes toward targets with which they implicitly identified.

To identify these contradictions, we examined the interaction between participants' gut feelings and each of the three indices of identity centrality described above. We use a similar analytic strategy across all of the resulting operationalizations—a pair of mixed linear models (one for A, one for B) that each include a random intercept for attitude domain and random slopes for identity centrality, gut feelings, and their interaction. Results of these analyses are summarized in Table 2.

Target identity centrality: When gut feelings contradicted attitudes toward identity-central targets, self-esteem was slightly lower. First, we examined target IDC. We predicted a positive coefficient for the interaction between gut feelings and target IDC, such that participants' negative reactions toward the attitude target would have a more negative effect on their self-esteem when that target is relatively central to their identity. For both A and B, the coefficient for the interaction term was positive and significantly different from 0 (b_A = 0.01, p = 0.001, 95% CI [0.005, 0.018], $N_{participants}$ = 3,714; b_B = 0.01[14], p = 0.020, 95% CI [0.001, 0.016], $N_{participants}$ = 3,716). See Table 6 and Figure 8. When analyzing just the confirmatory dataset instead of the merged dataset, the effects were less consistent. The interaction term was still positive for A (b_A = 0.01, p = .006), but not B (b_B = 0.01, p = 0.202). Overall, these results suggest that participants evaluated themselves less positively (by about half a point on the 6-point self-esteem scale) when they had very negative (compared to very positive) gut feelings about targets with which they directly identified (e.g., when participants reported that Asians, Artists, or Atheism were "part of" their self-concept).

Attitude identity centrality: When gut feelings contradicted identity-central attitudes, self-esteem was mostly unaffected. We next examined attitude IDC—the extent to which participants viewed accepting or rejecting A or B as important to or inconsistent with their self-concept. We found weak and inconsistent evidence of an interaction between gut feelings and attitude IDC (b_A = 0.004, p = 0.046, 95% CI [-0.000 , 0.009], $N_{participants}$ = 8,534; b_B = 0.004, p = 0.080, 95% CI [-0.000, 0.008], $N_{participants}$ = 8,549). See Table 7 and Figure 9. When analyzing just the confirmatory dataset instead of the merged dataset, the effects were even weaker. The interaction term was null for both A (b_A = .004, p = .098), and B (b_B = 0.004, p = 0.10). Overall, these results suggest that participants may have evaluated themselves very slightly less positively when they identified strongly as someone who has accepts (or rejects) something but nevertheless has very negative (or positive) gut feelings toward

14 This model did not converge when we attempted to estimate random slopes across participants, so it included only random intercepts.

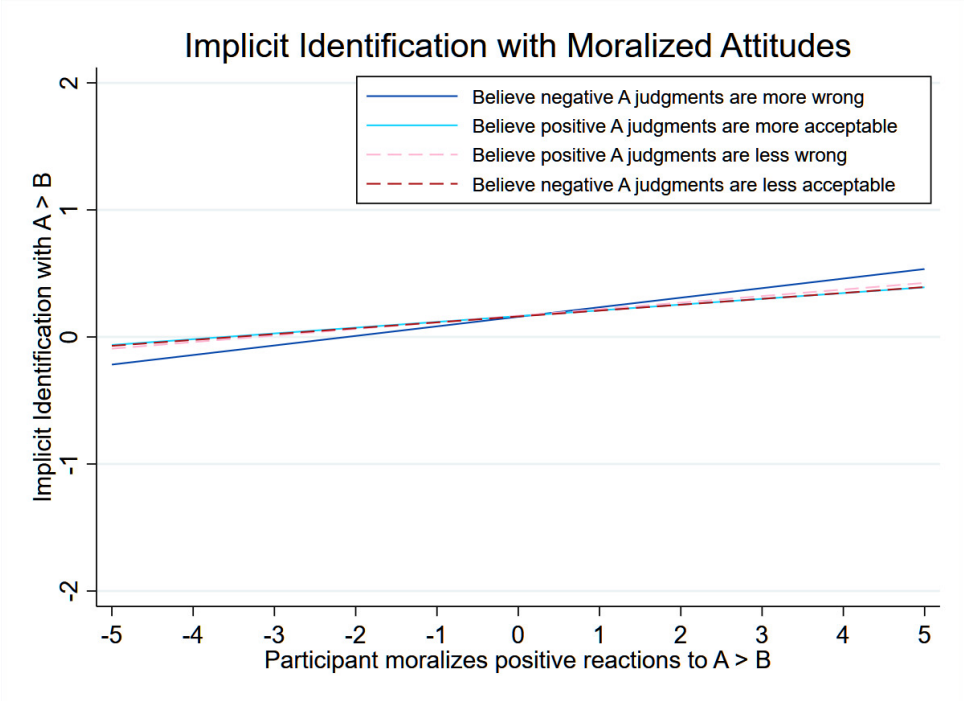

Figure 7. Lines indicate the predicted values of implicit identification across the full range of each of the four moralization variables. These predictions derive from the fixed portion of four separate linear mixed models, each of which predicted implicit identification from a single moralized attitude difference score, with random slopes and intercepts that varied across attitude pairs. Every difference score was computed such that higher values indicate moralizing more positive (or less negative) reactions to A relative to B.

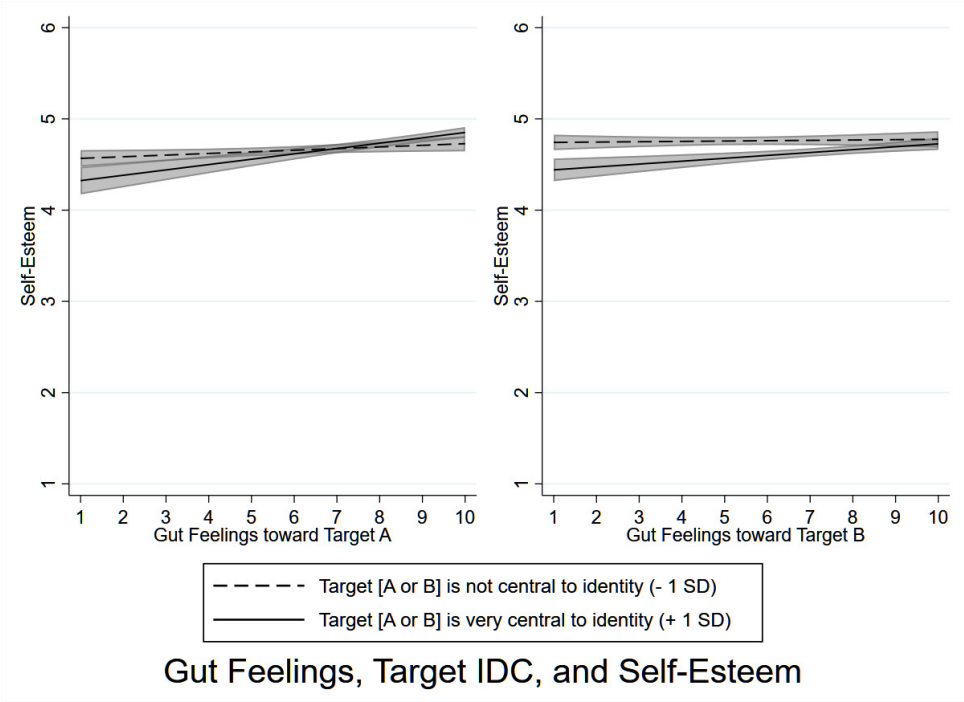

Figure 8. Lines indicate the predicted values of self-esteem across the full range of participants' gut feelings. Error bands indicate 95% CIs. Predictions derive from the fixed portion of two separate linear mixed models, described in Table 6.

Table 6. Predicting Self-Esteem from Gut Feelings about Identity Central Attitude Targets

Outcome: Self-Esteem	Attitude Target A		Attitude Target B	
	b (se)	*p*	*b* (se)	*p*
Fixed Effects				
Target Identity Centrality	-0.08**	0.020	-0.10***	<0.001
	(0.03)		(0.03)	
Gut Feelings about Target	-0.00	0.944	-0.01	0.55
	(0.01)		(0.01)	
Target IDC x Gut Feelings	0.01*	0.001	0.01*	0.020
	(0.00)		(0.00)	
Constant	4.68***	<0.001	4.87***	<0.001
	(0.08)		(0.08)	
Random Effects	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across attitude pairs)	0.00	(0.00)	<0.01	(0.00)
σ^2 (Random slope for Target IDC)	<0.01	(0.00)	(-)	(-)
σ^2 (Random slope for Gut Feelings)	<0.01	(-)	(-)	(-)
σ^2 (Random slope for T IDC x GF)	<0.01	(-)	(-)	(-)
Log-likelihood	-5047		-5,056	
Wald χ^2 (degrees freedom)	42.46* (3)		23.73** (3)	
N Level-1 units (participants)	3,714		3,716	
N Level-2 units (attitude pairs)	95		95	

Estimates derive from the two multi-level linear models depicted in Figure 8. The model for Target A included random slopes and intercepts that varied across attitude stimuli, whereas the model for Target B included only random intercepts that varied across stimuli. Stata 17 was sometimes unable to estimate standard errors for the variance in slopes and intercepts across stimuli, most likely because of the near-0 estimated variance in parameters across stimuli.
(†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

that thing. However, the magnitude of that effect (even when "significant" in one of these two models) was significantly smaller than .02, the smallest effect size of theoretical interest that we specified in our analysis plan (95% CI for Target A, 0.000, 0.009; For Target B, -0.000, 0.008).

Implicit identification: Self-esteem did not depend on implicit identities. Finally, we examined participants' implicit identification with attitude target A over attitude target B. We created a difference score indicating more favorable gut feelings toward A than toward B and interacted this variable with implicit identification in a mixed linear model like those above. The interaction was non-significant (b = 0.016, p = 0.125, 95% CI [-0.004, 0.037], $N_{participants}$ = 2,085; $N_{domains}$ = 95) and the 95% CI for the estimate only narrowly included our minimum threshold for a meaningful effect size (i.e., 0.0359). See Figure 10.

Pre-registered tests of alternative "rubrics." Our Identity Rubric hypothesis proposes that people evaluate themselves more negatively to the extent that their spontaneous affective reactions contradict attitudes central to their identity—that they would evaluate themselves based on whether they were "living up to" their identity-central attitudes. However, identity-central attitudes also tend to be more important, extreme, certain, and – as we have shown – connected to individuals' personal values (i.e., moralized). We therefore pre-registered analyses to distinguish living up to identity-central attitudes from living up to attitudes that were moralized, important, certain, or extreme.

We estimated several models that are summarized in Table 9. Our focus in each case was on the highest-order interaction term, indicating the extent to which attitude-inconsistent feelings were more or less predictive of participants' self-esteem. Above, we found that participants' gut feelings and identity-central attitudes often diverge without any relation to self-esteem. We find that these inconsistencies are also unrelated to self-esteem when attitudes

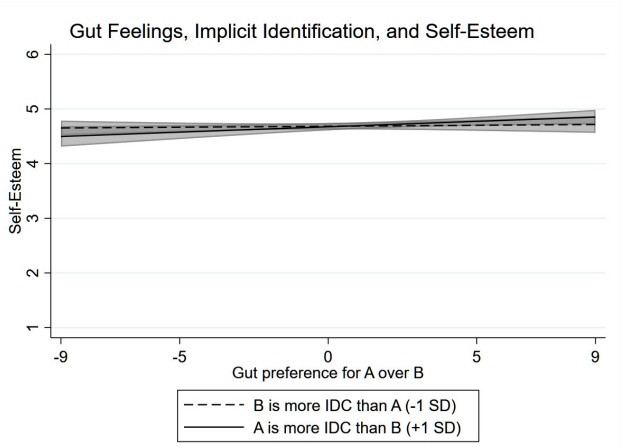

Figure 10. Lines indicate the predicted values of self-esteem across the full range of participants' gut preferences for A over B. Error bands indicate 95% CIs. Predictions derive from the fixed portion of a linear mixed model that included a random intercept permitted to vary across attitude pairs.

Table 7. Predicting Self-Esteem from Gut Feelings and Identity Central Attitudes

Outcome: Self-Esteem	Attitude Target A		Attitude Target B	
	b (se)	*p*	*b* (se)	*p*
Fixed Effects				
Attitude Identity Centrality	-0.01	0.472	-0.01	0.643
	(0.02)		(0.01)	
Gut Feelings about Target	0.004	0.691	-0.01	0.484
	(0.01)		(0.01)	
Attitude IDC x Gut Feelings	0.004*	0.046	0.004†	0.080
	(0.002)		(0.002)	
Constant	4.58***	<0.001	4.66***	<0.001
	(0.06)		0.06	
Random Effects	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across attitude pairs)	<0.01	(<0.001)	<0.01	(<0.001)
σ^2 (Random slope for Attitude IDC)	<0.01	(<0.001)	<0.01	(<0.001)
σ^2 (Random slope for Gut Feelings)	<0.01	(-)	<0.01	(<0.001)
σ^2 (Random slope for A IDC x GF)	<0.01	(<0.001)	<0.01	(.003)
Log-likelihood	-11,643		-11,672	
Wald χ^2 (degrees freedom)	46.18*** (3)		16.88*** (3)	
N Level-1 units (participants)	8,534		8,549	
N Level-2 units (attitude pairs)	95		95	

Estimates derive from the two multi-level linear models depicted in Figure 9, each of which included random slopes and intercepts that varied across attitude stimuli. Stata 17 was sometimes unable to estimate standard errors for the variance in slopes and intercepts across stimuli, most likely because of the near-0 estimated variance in parameters across stimuli.
(†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

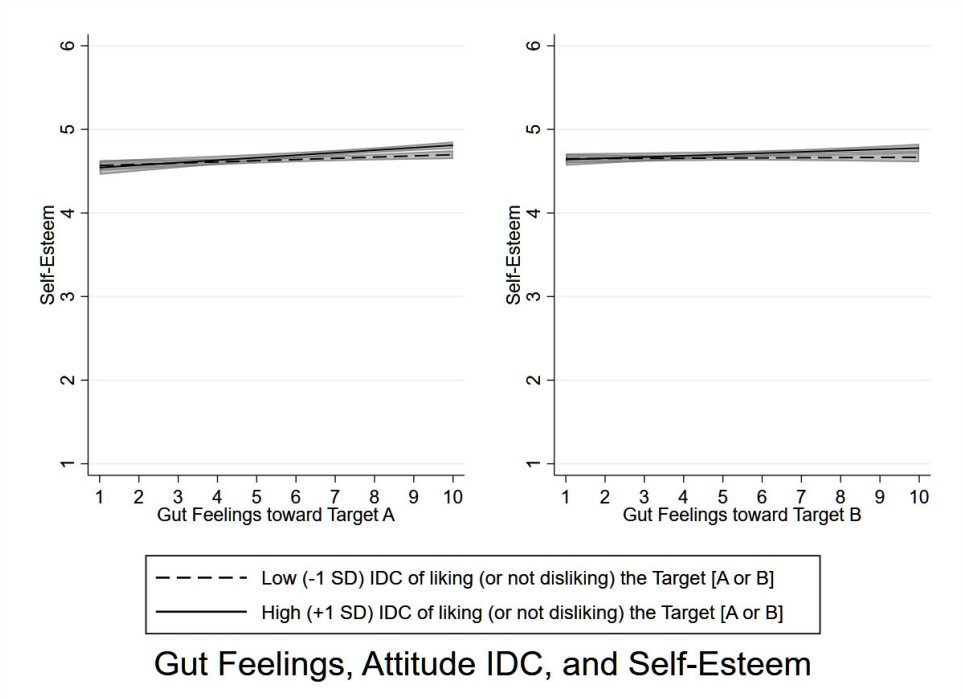

Figure 9. Lines indicate the predicted values of self-esteem across the full range of participants' gut feelings. Error bands indicate 95% CIs. Predictions derive from the fixed portion of two separate linear mixed models, each of which included random slopes and intercepts that varied across attitude pairs.

Table 8. Predicting Self-Esteem from Gut Preferences and Implicit Identification with Attitude Target

Outcome: Self-Esteem		
	b (se)	*p*
Implicit Identification with A over B	-0.009	0.836
	(0.044)	
Gut Preference for A over B	0.009	0.140
	(0.006)	
Implicit ID x Gut Preference	0.016	0.125
	(0.010)	
Constant	4.68***	<0.001
	(0.02)	
Random Effects	σ^2	(se)
σ^2 (Intercept across attitude pairs)	<0.01	(<0.001)
σ^2 (Random slope for Implicit ID)	<0.01	(<0.001)
σ^2 (Random slope for Gut Preference)	<0.01	(<0.001)
σ^2 (Random slope for I ID x Gut Pref)	<0.01	(<0.001)
Log-likelihood	-2,893	
Wald χ^2 (degrees freedom)	5.88 (3)	
N Level-1 units (participants)	2,085	
N Level-2 units (attitude pairs)	95	

Estimates derive from the multi-level linear model depicted in Figure 10, which included random slopes and intercepts that varied across attitude pairs. (†p < 0.10. *p < 0.05. **p < 0.01. ***p < 0.001.)

are moralized, important, extreme, or certain. The critical interaction was non-significant in 21 of the 24 models predicting self-esteem from moralized attitudes,[15] in all 6 models predicting self-esteem from attitude importance, in 4 of the 6 models predicting self-esteem from attitude certainty,[16] and in all 6 models predicting self-esteem from attitude extremity.

Discussion

Summary of Results

Our analyses yielded three key results. First, our descriptive analyses suggest that the attitudes that people perceive to be matters of right and wrong are extraordinarily diverse. Some participants saw room for debate concerning attitudes and behavior that most people would consider to be unequivocally moral. For example, a handful of participants "agreed" or "strongly agreed" that "Because of my personal values, I believe that making negative judgments about giving is acceptable" (6.3% of those who saw the question). At the same time, some participants passed stark judgment on some attitudes and behavior that most people would perceive as matters of taste. For example, some participants "strongly agreed" that "Because of my personal values, I believe that making negative judgments about Harry Potter is unacceptable" (8.5% of participants who saw the question). The descriptive statistics presented in Figure 4 illustrate this diversity. The end result is that there are probably very few (if any) attitudes or behaviors that all people would agree to be moral, immoral, or morally neutral. This diversity is in its own right an interesting characteristic of moral psychology, but it also suggests that researchers interested in how people think and feel about morally charged stimuli cannot always safely assume that their participants will construe a given stimulus in moral (or non-moral) terms.

Second, our results yield robust support for the Morality-Identity hypothesis. We found that when participants perceived their attitudes to be connected to their personal values, they were more likely to identify with those attitudes. They were also more likely to identify with the targets of those attitudes implicitly and explicitly (e.g., to see Harry Potter or Christians as part of their self-concept rather than to merely identify as someone who likes Harry Potter or Christians). This evidence is consistent with our prediction that people would perceive their idiosyncratic beliefs about right and wrong as a defining feature of who they are.

Finally, our results yield weak and inconsistent support for the Identity Rubric hypothesis. We find that our partici-

15 The few exceptions were in the models predicting self-esteem from Item #2 (the belief that positive judgments are wrong) and attitudes/feelings toward Target B.

16 The exceptions were only present in models focused on responses to Target B and ran in a counter-intuitive direction, such that participants reported lower self-esteem when their gut feelings were inconsistent with relatively *uncertain* attitudes.

Table 9. Tests of Alternative "Rubrics" for Self-Esteem

Rubric: Moralized Attitudes	Rubric: Important Attitudes	Rubric: Certain Attitudes	Rubric: Extreme Attitudes
Predictors in First Models	Predictors in First Models	Predictors in First Models	Predictors in First Models
Moralized Attitudes (Item 1, 2, 3, or 4)	Attitude Importance	Attitude Certainty	Attitude Extremity
Gut Feelings about Target	Gut Feelings about Target	Gut Feelings about Target	Gut Feelings about Target
	Actual Feelings about Target	Actual Feelings about Target	Actual Feelings about Target
	Gut Feelings X Actual Feelings	Gut Feelings X Actual Feelings	Gut Feelings X Actual Feelings
Moralized Atts x Gut Feelings	Importance X Gut Feelings	Certainty X Gut Feelings	Extremity X Gut Feelings
	Importance X Actual Feelings	Certainty X Actual Feelings	Extremity X Actual Feelings
	Importance X Gut X Actual	Certainty X Gut X Actual	Extremity X Gut X Actual
Predictors in Second Model	Predictors in Second Model	Predictors in Second Model	Predictors in Second Model
[Predictors in Initial Model]	[Predictors in Initial Model]	[Predictors in Initial Model]	[Predictors in Initial Model]
Target Identity Centrality	Target Identity Centrality	Target Identity Centrality	Target Identity Centrality
Target IDC X Gut Feelings	Target IDC X Gut Feelings	Target IDC X Gut Feelings	Target IDC X Gut Feelings
Predictors in Third Model	Predictors in Third Model	Predictors in Third Model	Predictors in Third Model
[Predictors in Initial Model]	[Predictors in Initial Model]	[Predictors in Initial Model]	[Predictors in Initial Model]
Attitude Identity Centrality	Attitude Identity Centrality	Attitude Identity Centrality	Attitude Identity Centrality
Attitude IDC X Gut Feelings	Attitude IDC X Gut Feelings	Attitude IDC X Gut Feelings	Attitude IDC X Gut Feelings

Note. Entries indicate the predictors we used to investigate how participants evaluated themselves when their spontaneous affective reactions contradicted attitudes that were important, extreme, certain, or moralized (vs. identity-central). The outcome in every model is participants' self-esteem. All models allowed slopes and intercepts to vary randomly across attitude stimuli (unless models failed to converge, in which case we only allowed the intercept to vary).
"Actual feelings" were omitted from models including the moralized attitude measure because unlike the AIID measures of importance, certainty, and extremity, "actual feelings" are effectively "baked in" to the AIID measures of moralized attitudes and identity centrality.

pants did not consistently evaluate themselves on the basis of whether their gut feelings were consistent with the "actual" attitudes that they cherish as defining features of their identity. Our first analysis found some evidence suggestive of this phenomenon; when participants reported negative gut feelings about targets with which they identified, they also reported slightly lower self-esteem. However, this may be because negative gut feelings about identity-central targets are, in and of themselves, negative gut feelings *about the self*. For example, negative gut feelings about Christians, Jews, European Americans, and African Americans could easily translate to negative self-evaluations among people who are themselves Christian, Jewish, European American, or African American—regardless of whether they think it is morally "wrong" or "acceptable" to judge people from these groups positively. Other models were more directly inconsistent with our predictions. When participants reported negative gut feelings about targets that they thought they *ought* to like, their reported self-esteem was almost identical to that of participants whose gut feelings and explicit attitudes were perfectly consistent. We observed a similar null effect for implicit identification with targets and for every other indicator of attitude strength that we analyzed. Participants generally did not have lower self-esteem when their gut feelings were inconsistent with moralized, important, certain, or extreme attitudes.

Based on these findings, our initial theory requires significant revision. We hypothesized that moralized attitudes inform self-evaluation because they structure individuals' self-concepts. Although we did find that moralized attitudes were relatively central to participants' identities (consistent with the Morality-Identity hypothesis), we found little if any connection between participants' self-esteem and the extent to which their gut feelings were consistent with those attitudes. Participants did not report feeling meaningfully worse about themselves when they were attracted to things they believed they shouldn't like or repulsed by things they believed they ought to accept.

Limitations

That said, our study has several limitations that complicate this test of our theory.

Mixed evidence for key measures' validity. The evidence from our validity study was mixed and weaker than we had anticipated. The most widely used measure of attitude moralization (moral conviction) only sometimes predicted responses to the AIID items. For one item, the relation was what we predicted for both positive and negative attitudes, though those relations were weak. For the other three items, the relation was only present either for positive or negative attitudes. We encountered similar problems with the AIID measures of attitude identity centrality. Despite this limitation, we remain confident in our conclusions for three reasons.

First, despite evidence that some measures were more valid than others and more or less valid for positive versus negative attitudes, we find no evidence that our results de-

pended on which AIID items we used to assess attitude moralization or attitude identity centrality. Second, although the measures we used to assess the validity of the AIID items are certainly more widely used and better established as measures of the target constructs, they are still just measures, not perfect reflections of the constructs of interest. Given that the AIID measures each make explicit reference to their respective target constructs (e.g., "personal values," decisions about what is "wrong" or "acceptable," whether reactions are "important to" and "inconsistent with" the "self-concept"), these measures may capture aspects of moralization and identity centrality that other measures do not. Finally, we found converging evidence for our conclusions with analyses that does not rely on the AIID measures. To do this, we used the validation study to conduct an additional un-pre-registered test of the Morality-Identity hypothesis. Our theory would predict that Skitka and colleagues' moral conviction measure would be related to Luhtanen and Crocker's (1992) measure of identity centrality for the 20 attitude targets in this study. It was (b = 0.32, 95% CI: [0.27, 0.37], $p < 0.001$), even controlling for the effect of importance. In addition, the analyses described in Table 9 suggest that alternative measures of moralization or identification would probably not yield any stronger support for the Identity Rubric hypothesis. Participants' self-esteem was basically unmoved regardless of how severely their gut feelings contradicted certain, important, extreme, "identity-central," or "moralized" attitudes. Even if we have failed to find a direct, precise measure of moralization or identity centrality, surely at least one of these indices of attitude strength would at least be correlated with such a measure. If the Identity Rubric hypothesis were true, then, it seems unlikely that all of these tests would be so uniformly null.

Still, our validity study offers an important caveat for our own and future work. We cannot safely assume that when people say that their attitude is moral or important to who they are, they will *also* say that the opposite attitude is immoral or anathema to their self-concept. Although we might say people "moralize" an attitude when they judge it to be desirable, acceptable, or wrong, these judgments probably do not lie along a single dimension.

Narrow threats to the self and a broad measure of self-esteem. In hindsight, our predictions may have presumed self-esteem to be too fragile. The Rosenberg Self-Esteem scale is intended to measure self-esteem as a global trait. Meanwhile, we analyzed participants' gut feelings and attitudes toward only one or two targets. Contradictions so narrow and specific may be insufficient to impact global trait-level self-esteem in a meaningful way.

On the one hand, narrower measures of self-esteem might prove to be more malleable. People's thoughts and behavior during a particular event or time period might affect how they feel about themselves during that specific time. On the other hand, more frequent, numerous, or chronically salient contradictions between people's gut feelings and the attitudes they believe to be appropriate might have a stronger impact on self-esteem than the one or two attitudes we were able to assess.

Some scholars have conducted experiments to confront participants with moral failures and trace their impact on self-evaluations (see Wojciszke, 2005). These participants often end these experiments feeling just fine. These results are consistent with decades of social-psychological research have documented individuals' capacity to rationalize their behaviors and what they might consider to be failures to practice what they preach, and scholars have often argued that the point of this rationalization is to protect people's positive self-images (e.g., Aronson, 1969; Kunda, 1990).

Small wonder, then, that slight divergences between "gut feelings" and "actual" attitudes in a single attitude domain failed to leave a dent in our participants' self-esteem. Our study leaves open the possibility that some attitude-inconsistent feelings or behavior may be more uncomfortable than others, and that measures of more specific or shorter-term self-evaluation might be more likely to change in the wake of these behaviors. Future work might find that when people behave in ways that are clearly at odds with multiple important, moralized, or identity-central attitudes, they briefly feel worse about themselves.

Correlational Design. We have tested a multi-step causal framework with a correlational dataset, which cannot permit strong causal inferences. For example, we cannot know whether people come to identify with certain attitudes because they see them in moral terms, moralize attitudes because they are central to their identity, or come to moralize and identify with their attitudes simultaneously as a part of some larger process. The reality is probably a combination of these possibilities. For example, someone may come to identify with their pro-choice attitude because they see abortion access as a moral issue and *also* come to moralize their attitude toward Britney Spears, 50 Cent, or Harry Potter because they identify as a fan. Regardless, our evidence suggests that these processes are connected. At the same time, the latter part of the model proposed in Figure 1 is now less plausible, as our correlational evidence was inconsistent with the Identity Rubric hypothesis.

Conclusion

Our goal with the present study was to clarify the nature and function of moralized attitudes and their role in self-evaluation. Existing research indicates that people evaluate others based on their morality and that people are motivated to preserve a moral self-image, but it is less clear whether people evaluate themselves positively or negatively based on to the extent to which they embody the idiosyncratic moral self-image to which they aspire. We directly tested this possibility. We find that although people see their moralized attitudes as a defining feature of who they are, that does not mean that they evaluate themselves more negatively when their spontaneous affective reactions are inconsistent with those attitudes.

At the same time, our findings reveal some uncertainties that future work might investigate. First, our validation study suggests that beliefs about what is "acceptable" do not precisely mirror beliefs about what is "wrong," and we did not measure or analyze any data about what participants believe to be "right" or morally desirable. Future efforts investigating how and whether people moralize their attitudes could explicitly examine the processes by which people moralize positive and negative attitudes and by

which people come to decide that either type of attitude is right, wrong, or merely acceptable. Second, our investigation leaves the possibility that attitude-inconsistent feelings *could* impact self-esteem in ways that we might observe if we considered shorter-term fluctuations in self-esteem or instances in which individuals' feelings (or behaviors!) betrayed their ostensible values more strongly, more frequently, or more blatantly. Finally, our cross-sectional study cannot reveal how people come to develop their moral beliefs or come to identify with their attitudes over time. Longitudinal or experimental work that examines a larger variety of attitudes *within* individuals and measures more dynamic and specific fluctuations in self-evaluation could help advance our knowledge in these areas.

For now, however, our findings suggest that trait-level self-esteem is relatively resilient to attitude-inconsistent gut feelings. Even in identity-central domains, people may rationalize or ignore their attitude-inconsistent feelings to protect their self-esteem. We did not find evidence that people feel worse about themselves when they fall short of who they think they ought to be.

Contributions

Dr. P.D. Ekstrom conceived the research with assistance from Dr. C.K. Lai. Dr. Ekstrom and Dr. Lai worked together to design the research and analysis plan. Dr. Ekstrom analyzed the data. Dr. Ekstrom and Dr. Lai worked together to interpret the data. Dr. Ekstrom wrote the paper, and he and Dr. Lai revised the paper together.

Competing Interests

Neither Dr. Ekstrom nor Dr. Lai have any competing interests to declare. Dr. Lai was one of the co-authors on the AIID dataset, but did not access the data prior to our analysis or perform any analyses. Dr. Lai is also the Chair of the Scientific Advisory Board and a consultant with Project Implicit, a non-profit organization. Project Implicit's mission is to support research on implicit social cognition and educate the public about bias.

Data Accessibility Statement

All the survey questions, task materials, participant data, and analysis scripts are available at OSF. For the original AIID dataset, see: https://osf.io/pcjwf/. For all of our analyses using these data, see: https://osf.io/6ckns/. Our Stage 1 registered report (which includes our analysis plan and exploratory results) is also available at the latter OSF link, and registered here: https://doi.org/10.17605/osf.io/zry5b.

Submitted: December 22, 2021 PDT, Accepted: May 23, 2022 PDT

References

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology* (Vol. 4, pp. 1–34). Academic Press. https://doi.org/10.1016/s0065-2601(08)60075-1

Boninger, D. S., Krosnick, J. A., Berent, M. K., & Fabrigar, L. R. (1995). The causes and consequences of attitude importance. In R. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 159–189).

Ebersole, C. R., Hussey, I., Hughes, S., Axt, J., Lai, C. K., & Nosek, B. A. (2019). The Attitudes, Identities, and Individual Differences (AIID) Study and Dataset. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/PCJWF

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). Academic Press.

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition*, *25*(5), 687–717. https://doi.org/10.1521/soco.2007.25.5.687

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Graham, J., Meindl, P., & Beall, E. (2012). Integrating the streams of morality research: The case of political ideology. *Current Directions in Psychological Science*, *21*(6), 373–377. https://doi.org/10.1177/0963721412456842

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*(1), 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.146

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General*, *134*(3), 420–425. https://doi.org/10.1037/0096-3445.134.3.420

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Janoff-Bulman, R., Sheikh, S., & Baldacci, K. G. (2008). Mapping moral motives: Approach, avoidance, and political orientation. *Journal of Experimental Social Psychology*, *44*(4), 1091–1099. https://doi.org/10.1016/j.jesp.2007.11.003

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology*, *110*(5), 660–674. https://doi.org/10.1037/pspa0000050

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. *Attitude Strength: Antecedents and Consequences*, *1*, 1–24.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272–1290. https://doi.org/10.1177/0146167216655984

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin*, *18*(3), 302–318. https://doi.org/10.1177/0146167292183006

McFerran, B., Aquino, K., & Duffy, M. (2010). How personality and moral identity relate to individuals' ethical ideology. *Business Ethics Quarterly*, *20*(1), 35–56. https://doi.org/10.5840/beq20102014

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344–357. https://doi.org/10.1111/j.1751-9004.2010.00263.x

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Prentice, M., Jayawickreme, E., Hawkins, A., Hartley, A., Furr, R. M., & Fleeson, W. (2019). Morality as a Basic Psychological Need. *Social Psychological and Personality Science*, *10*(4), 449–460. https://doi.org/10.1177/1948550618772011

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*(2), 386–396. https://doi.org/10.1016/j.jesp.2006.12.008

Reed, A., II, & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of Personality and Social Psychology*, *84*(6), 1270–1286. https://doi.org/10.1037/0022-3514.84.6.1270

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press. https://doi.org/10.1515/9781400876136

Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, *53*(3), 550–562. https://doi.org/10.1037/0022-3514.53.3.550

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281. https://doi.org/10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, *88*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895

Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. In H. Lavine (Ed.), *Political Psychology* (Vol. 35, pp. 95–110). Wiley. https://doi.org/10.1111/pops.12166

Stanley, M. L., Henne, P., Iyengar, V., Sinnott-Armstrong, W., & De Brigard, F. (2017). I'm not the person I used to be: The self and autobiographical memories of immoral actions. *Journal of Experimental Psychology: General*, *146*(6), 884–895. https://doi.org/10.1037/xge0000317

StataCorp. (2021). *Stata Statistical Software: Release 17*. StataCorp LLC.

Tenbrunsel, A. E., Diekmann, K. A., Wade-Benzoni, K. A., & Bazerman, M. H. (2010). The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are. *Research in Organizational Behavior*, *30*, 153–173. https://doi.org/10.1016/j.riob.2010.08.004

Visser, P. S., Bizer, G. Y., & Krosnick, J. A. (2006). Exploring the latent structure of strength-related attitude attributes. *Advances in Experimental Social Psychology*, *38*, 1–67. https://doi.org/10.1016/s0065-2601(06)38001-x

Winterich, K. P., Mittal, V., & Ross, W. T., Jr. (2009). Donation behavior toward in-groups and out-groups: The role of gender and moral identity. *Journal of Consumer Research*, *36*(2), 199–214. https://doi.org/10.1086/596720

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232. https://doi.org/10.1037/0022-3514.67.2.222

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A. E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology*, *41*(5), 617–627. https://doi.org/10.1002/ejsp.791

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Wright, J. C., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality & Social Psychology Bulletin*, *34*(11), 1461–1476. https://doi.org/10.1177/0146167208322557

Appendix A: Validation Study Questionnaire

Implicit Identification IAT

We will use the same verbal and visual stimuli and the same scoring procedure as the original AIID dataset for all stimulus pairs. See https://osf.io/pcjwf/ for documentation and see below for possible stimulus pairs.

Explicit Attitudes

How positive or negative do you feel towards [A/B]?

Response options:

1 – Strongly negative, 2, 3, 4, 5, 6, 7, 8, 9, 10 – Strongly positive

AIID Moralized Attitudes

Because of my personal values, I believe that making negative judgments about [A/B] is wrong.

Because of my personal values, I believe that making positive judgments about [A/B] is wrong. (reverse-scored)

Because of my personal values, I believe that making negative judgments about [A/B] is acceptable. (reverse-scored)

Because of my personal values, I believe that making positive judgments about [A/B] is acceptable.

Response options: 1 = Strongly disagree 2 = Disagree 3 = Slightly disagree 4 = Slightly agree 5 = Agree 6 = Strongly agree

AIID Attitude Identity Centrality

Being rejecting of A/B is inconsistent with my self-concept.

Being accepting of A/B is inconsistent with my self-concept. (reverse-scored)

*Being rejecting of A/B is consistent with my self-concept. (reverse-scored)

*Being accepting of A/B is consistent with my self-concept.

Response options: 1 = Strongly disagree 2 = Disagree 3 = Slightly disagree 4 = Slightly agree 5 = Agree 6 = Strongly agree

**Due to a transcription error on our part, two attitude IDC items differed from their original wording in the AIID questionnaire. In our validation study, Items 3 and 4 asked participants whether rejecting or accepting A/B was "consistent with [their] self-concept." The original AIID asked whether rejecting or accepting A/B was "important to [their] self-concept."*

AIID Target Identity Centrality

How much is A/B part of your self-concept?

1 = None at all [sic; this was the response label originally used in the AIID], 2, 3, 4, 5,

6 = Very much

Standard Measure of Attitude Moralization (Moral Conviction; Skitka & Morgan, 2014)

To what extent are your feelings about [A/B] a reflection of your core moral beliefs and convictions?

To what extent are your feelings about [A/B] connected to your beliefs about fundamental right and wrong?

To what extent are your feelings about [A/B] based on moral principle?

To what extent are your feelings about [A/B] a moral stance?

Response options: Not at all, Slightly, Moderately, Much, and Very much

Standard Measure of Attitude Identity Centrality (modified from Luhtanen & Crocker, 1992)

Overall, my feelings about [target] have very little to do with how I feel about myself.

My feelings about [target] are an important reflection of who I am.

My feelings about [target] are important to my sense of what kind of person I am.

Response options: 1 = Strongly disagree 2 = Disagree 3 = Slightly disagree 4 = Slightly agree 5 = Agree 6 = Strongly agree

Attitude Importance (modified from Boninger et al., 1995)

How important to you are your feelings about [A/B]?

Response options: Not at all important, Slightly important, Somewhat important, Important, Very important

How deeply do you care about [A/B]?

Response options: Not at all, Slightly, Somewhat, Quite a bit, Very deeply

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/36344-a-good-person-shouldn-t-feel-this-way-moralized-attitudes-identity-and-self-esteem/attachment/91993.docx?auth_token=G1iW7tbDOTLHbdjEZR12

Supplementary Analyses

Download: https://collabra.scholasticahq.com/article/36344-a-good-person-shouldn-t-feel-this-way-moralized-attitudes-identity-and-self-esteem/attachment/91994.docx?auth_token=G1iW7tbDOTLHbdjEZR12