A good person shouldn't feel this way:
Moralized attitudes, identity, and self-esteem

Pierce Ekstrom
University of Nebraska – Lincoln

Calvin Lai
Washington University in St. Louis

Registered report accepted in principle at *Collabra: Psychology* on July 2nd, 2021

Abstract

Moralized attitudes are the attitudes that people construe as matters of right and wrong. In this study, we examine how moralized attitudes influence how people perceive and 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 exploratory analyses of a subset of the AIID dataset, 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 found tentative evidence that participants reported lowered feelings of self-worth when they experienced mental conflict between attitudes that were central to their identity and their gut feelings toward the objects of those attitudes. With future access to the remainder of the AIID dataset, we will repeat these analyses in confirmatory tests of our hypotheses. Our findings will clarify the role that morality plays in self-perception and whether people think less of themselves when they fall short of the people they aspire to be.

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

A good person shouldn't feel this way:
Moralized attitudes, identity, and self-esteem

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, Cuddy, & Glick, 2007; Goodwin, Piazza, & Rozin, 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' (2005; Skitka, 2010) 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 will first test whether moralized attitudes are more central to individuals' identities than non-moral attitudes. Second, we will test whether identity-central attitudes provide a rubric for self-evaluation. When people's spontaneous feelings about a target contradict what they believe their "real" attitude to be, we expect participants to report lower

self-esteem to the extent that the attitude is central to their identity. In sum, we propose that individuals' moralized attitudes help define their identities, which 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, Cuddy, & Glick, 2007; Wojciszke, 2005; Wojciszke, 1994; Wojciszke, Bazinska, Jaworski, 1998) and sometimes even interpersonal warmth (Landy, Piazza, & Goodwin, 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, Mullen, & Murnighan, 2011; Klein & Epley, 2016; Prentice, Jayawickreme, Hawkins, Hartley, Furr, & Fleeson, 2019; Stanley, Henne, Iyengar, Sinnott-Armstrong & Brigard, 2017; Merritt, Effron, & Monin, 2010; Tenbrunsel, Diekmann, Wade-Benzoni, & Bazerman, 2010), suggesting that morality is a valued component of the self and an impactful component of identity (Aquino & Reed, 2002; McFerran, Aquino, & Duffy, 2010; Reed & Aquino, 2003; Winterich, Mittal, & Ross, 2009; Wojciszke, Baryla, Parzuchowski, Szymkow, & Abele, 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 (Janoff-Bulman, Sheikh, & Baldacci, 2008; Haidt, 2012; 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, Bauman, & Sargis, 2005) and more willingness to openly discriminate against moral dissenters compared to those who disagree with them in non-moral domains (Wright, Cullum, & Schwab, 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 examine whether the idiosyncratic attitudes that individuals moralize 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 argue that any attitude imbued with moral significance can 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 tell you, 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, Smith, & Nosek, 2008).

We predict 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 s/he moralizes and thus identifies 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.

Our 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, 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 “attitude strength” dimension (Visser, Bizer, & Krosnick, 2006), so any one may be of more or less interest in a given theory. 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 predict that moralized attitudes will be uniquely central to individuals' identities, above and beyond how certain, important, or extreme they are. We also predict that when spontaneous affective reactions and self-reported attitudes are inconsistent, people will report lower self-esteem to the extent that attitudes are central to their identity—again, above and beyond how certain, important, or extreme their attitudes are. Certain, important, and extreme attitudes may be central to individuals' identities, and individuals may feel uncomfortable when

their spontaneous affective reactions contradict them, but we plan to 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) 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 begin by proposing a study to validate the measures employed in the AIID, some of which were designed for purposes different than our own.

We then present 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 that certain patterns of thought or behavior (e.g., generosity) are consistently moralized across individuals (see Graham, Meindl, & Beall, 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.

Finally, we propose to test two hypotheses that lay the foundation for our core theoretical argument. If moralized attitudes inform self-evaluation because they structure individuals' self-concepts, then at least two things must be true.

1. Moralized attitudes must be relatively central to identity (Morality-Identity hypothesis). We predict that moralized attitudes comprise a central component of individuals' self-concepts. These attitudes are part and parcel of how people see themselves—more so than non-moralized attitudes. We will test this hypothesis by predicting multiple measures of attitudes' identity centrality from the extent to which participants report their attitudes to be connected to their personal values. When participants claim their attitude is value-relevant, we expect them to identify relatively closely with that attitude and to identify with the target of that attitude. 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 predict that people feel worse about themselves when their spontaneous affective reactions are inconsistent with identity-central attitudes. We will test 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 their identity). We also expect these effects to hold above and beyond attitude importance, certainty, and extremity.

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 our theory to pre-registered tests. Evidence inconsistent with either the Morality-Identity hypothesis or the Identity-Rubric hypothesis would require a fundamental revision to our theory.

In sum, the current paper (1) proposes a novel study to validate measures employed in the AIID, (2) presents key descriptive statistics from the AIID, and (3) presents preliminary results from the hypothesis tests we will conduct assuming the AIID measures are at least partly valid indices of our constructs of interest. Specifically, we have tested the Morality-Identity hypothesis and the Identity Rubric hypothesis in a subset of the AIID data (about 15% of the full dataset) that its curators have made available to us to conduct exploratory analyses. We found strong support for the Morality-Identity hypothesis and inconsistent support for the Identity Rubric hypothesis. Pending the results of our validity study, we will repeat our tests of these hypotheses (using the same models) by *adding* the remaining 85% of the AIID dataset to our existing data and re-estimating our models using this complete dataset. We may refrain from re-estimating some models if our validation study suggests that predictors used in those models are not valid indicators of target constructs.

Figure 1. Illustration of hypotheses.

AIID Method

AIID Data

Our primary data come from the Attitudes, Identities, and Individual Differences (AIID) Study (see <https://osf.io/pcjwf/> for documentation). Participants (total $N \approx 200,000$) were randomly assigned to answer questions about one of 95 attitude pairs (“domains” in the AIID documentation; 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 probably 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, Westfall, & Kenny, 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% White, 6% Black, 6% Hispanic, 7% Asian, 1% Native American, 8% other or multi-racial; Age ranged from 7 to 88 years, $M = 30$, $SD = 11.78$). The sample was more educated and liberal than the general population. Of those who reported their education ($n = 46,654$ in the exploratory dataset), 86% claimed at least some college education, and 56% of those who reported their political identity ($n = 45,704$) 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 ($n = 36,975$), Canada ($n = 2,277$), and the United Kingdom ($n = 2,181$).² Data collection took place between 2004 and 2007.

We have performed an exploratory analysis using data from a random subset of about 15% of AIID participants. We will combine these data with the remaining 85% of the dataset to perform confirmatory tests of our hypotheses.

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

² Note that all of these *ns* are derived from the exploratory subset of the AIID data, which includes 15% of cases from the full dataset.

Table 1. Number of Observed Cases per Measure

Variable	Exploratory Subset <i>N</i>		Confirmatory Subset <i>N</i>	
Implicit Identification	7,640		38,405	
Self-Esteem	1,505		7,325	
	<u>Target A</u>	<u>Target B</u>	<u>Target A</u>	<u>Target B</u>
Gut feelings	31,472	31,482	152,318	152,362
Target IDC	13,550	13,554	65,005	64,992
Moralized Attitude Composite	31,168	31,174	150,660	150,627
Moralized Attitude Item 1	7,709	7,705	37,863	37,870
Moralized Attitude Item 2	7,843	7,843	37,665	37,657
Moralized Attitude Item 3	7,837	7,844	37,409	37,390
Moralized Attitude Item 4	7,779	7,782	37,723	37,710
Attitude IDC Composite	31,141	31,132	150,552	150,575
Attitude IDC Item 1	7,862	7,865	37,663	37,687
Attitude IDC Item 2	7,814	7,814	37,356	37,360
Attitude IDC Item 3	7,695	7,698	37,869	37,885
Attitude IDC Item 4	7,770	7,755	37,634	37,643

Note. We identified the number of cases in the confirmatory subset of the AIID using a “masked” version of the dataset, in which all non-missing quantitative participant data were replaced with a value of 1.

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 supplement).

Next, in random order, participants completed an Implicit Association Test (IAT) and a self-report questionnaire focused on those attitude targets (Greenwald, McGhee, & Schwartz, 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.”³ During the questionnaire, participants responded to a partly random⁴ 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., 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.

³ In this article, we use “implicit” to refer to indirect measurement. We do not make claims about the exact nature of the constructs measured by Implicit Association Tests.

⁴ See <https://osf.io/pcjwf/> for randomization details.

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.⁵ We expect these models to yield substantively identical results.

Gut 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). 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, Smith, & Nosek, 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

⁵ 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.

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 proposed validation study, we will assess whether participants' scores on these items are systematically related to the extent that they moralize the measured attitude.

To maximize our number of analyzable cases, some analyses use 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 will therefore assess the reliability of this composite in our proposed validation study. Pending those results, we currently examine the effects of each moralization variable and the composite separately (i.e., in separate regression models).

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*.

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

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 will use 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 ($N = 1,505$ in the exploratory dataset; about 5% of all cases) were assigned to complete the Rosenberg Self-Esteem Scale (1965). 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’ global evaluations of the attitude target. Global evaluations were measured using either positivity (“How positive or negative do you feel towards [A/B]?” from 1 “Strongly Negative” to 10 “Strongly Positive”), warmth (“How warm or cold do you feel towards [A/B]?” from 1 “Cold” to 10 “Warm”), or liking (“How much do you like or dislike [A/B]?” from 1 “Strongly Dislike” to 10 “Strongly Like”).

Proposed Validation Study

Because the AIID items measuring moralized attitudes and attitude identity centrality confound attitude valence with moralization and identification (respectively), we propose to collect new data to test how well the AIID items measure these core constructs independent of valence. These new data will 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 will be 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 will aim to recruit at least 800 participants. See supplemental document “Power Analyses ” for detailed information about the power analysis.

Validation Study Procedure

Each participant will be randomly assigned to answer questions about one pair of stimuli from the AIID. All participants will first complete an identification IAT for that stimulus pair, then report their explicit attitudes toward each stimulus. Next, participants will complete 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, Krosnick, Berent, & Fabrigar, 1995). Participants will complete each measure 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 will report “How positive or negative do you feel towards [A/B]?” on a scale from 1 to 10.

AIID Moralized attitudes. Each participant will complete 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 will complete 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).

AIID Target Identity Centrality. We will measure 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 will use 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 will use 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 will indicate the importance of their attitudes toward A and B using two items used by Boninger et al. (1995).

Attitude Target Stimuli. Participants will be 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.⁶

Validation Study Analysis Plan

Our validation study seeks to assess the internal consistency and validity of the AIID’s moralized attitude items and attitude identity centrality items. This study will also enable us to examine the correlates of target identity centrality and implicit identification.

⁶ 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).

Evaluating the AIID's Moralized Attitude Measures. To assess the internal consistency of the AIID's moralized attitude measures, we will compute a series of Cronbach's α s. We will compute a separate α across the four moralized attitude items for each target stimulus (e.g., Tall People, Short People, Jews, Christians, Strong). If the median of the resulting 20 α s is below 0.6, we will drop all proposed analyses using the "composite" measure of moralized attitudes. If the median α is greater than 0.6, we will retain the composite measure and planned analyses, with appropriate caveats if α s are low for specific stimulus pairs.

To assess the validity of the AIID's moralized attitude measures, we will estimate multi-level models predicting responses to each of the four items and their composite from moral conviction, attitude valence, and the interaction between these two predictors, controlling for importance and its interaction with attitude valence. This model will include 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 will allow the model intercept to vary randomly across participants and across stimulus pairs.

People who moralize their attitudes should perceive attitudes opposed to their own as more "wrong" and less "acceptable." These moralizers should also perceive attitudes in line with their own as especially "acceptable" and less "wrong" than non-moralizers.

Operationally, we plan to use the conditional effects (i.e., the simple slopes) of moral conviction to assess the validity of each moralized attitude item. For example, moral conviction should positively predict the belief that "negative judgments about Christians are wrong" among participants whose attitudes toward Christians are positive. Moral conviction should negatively predict the same belief among participants whose attitudes toward Christians are negative. The same logic applies to the other three items and to the composite. Because all models will control

for attitude importance and its interaction with valence, these simple slopes will tell us the extent to which the AIID moralized attitude items tap moralization *per se* (rather than just attitude strength in general).

The key conditional effects may vary in their size and significance both within and across models. For instance, moral conviction may be more closely tied to the belief that opposing attitudes are “wrong” than to the belief that one’s own attitude is “acceptable.” For this reason, we will refrain from binary determinations of validity or invalidity. Instead, we will discuss the *extent to which* each item is valid based on the significance *and* the size of the conditional effects of moral conviction upon it.

Evaluating the AIID’s Attitude Identity Centrality Measures. We will use the same basic strategies to evaluate the AIID’s identity centrality measures. To assess their internal consistency, we will again compute separate Cronbach’s α s for each of our 20 target stimuli. Once again, if the median of the resulting 20 α s is below 0.6, we will drop all proposed analyses using the “composite” measure.

To assess the validity of the AIID’s attitude identity centrality measures, we will estimate multi-level models predicting responses to each of the four items and their composite. Predictors will include Luhtanen and Crocker’s (1992) measure of identity centrality, attitude valence, and the interaction between these two predictors. We will not include other covariates. Once again, because this model will include two observations for each participant and many participants per stimulus pair, we will allow the model intercept to vary randomly across participants and across stimulus pairs.

Again, our assessment of each item’s validity will hinge on conditional effects. People who score 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. This logic implies a distinct pattern of conditional effects for each item. We will use both the significance and the magnitude of these conditional effects to infer the validity of each attitude identity centrality item.

Target Identity Centrality and Implicit Identification. Our validation study will also include the AIID’s measures of target identity centrality and implicit identification. Their correlations with the AIID attitude identity centrality items will provide secondary evidence of the latter’s validity.

Exploratory Results from AIID

Assuming that our validation study indicates that all key measures are sufficiently valid to be analytically useful, we have conducted preliminary tests of the Morality-Identity and Identity Rubric hypotheses using the available AIID data. All files necessary to reproduce our exploratory analyses are provided in our online supplemental materials, accessible at the following link: https://osf.io/6ckns/?view_only=794bdcfca57f40bb9ecfca0389c6debc. These files include the raw and processed data files, the scripts we used to conduct the analysis (in Stata 16), and a codebook that indicates which variables in the dataset correspond to those in the manuscript.

Moralization and Identity Centrality Across Domains and Targets

Before we examine the connection between attitude moralization and identity centrality at the individual level, we ask whether certain attitude targets are more frequently moralized or identity-central than others. 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.⁷ 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 2 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 at least 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, $b = 1.10$, $p < 0.001$, 95% CI [1.02, 1.17]; for target IDC, $b = 1.12$, $p < 0.001$, 95% CI [0.84, 1.41]).⁸ 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 ($M = 4.63$, $SD = 1.77$) being more consistently moralized than

⁷ The fact that any given AIID participant saw only one pair of attitude targets precludes cluster analysis and other methods that might support a more data-driven categorization scheme.

⁸ Regression coefficients derive from OLS regressions using a target-level dataset (i.e., one row for each unique attitude target).

positive reactions to abstract ideas ($M = 4.36$, $SD = 1.58$), everyday targets ($M = 4.16$, $SD = 1.77$), and political targets ($M = 4.02$, $SD = 1.72$).

A third important pattern within these means 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 4.86 ($SD = 1.38$). 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 3.10 ($SD = 1.84$). 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 ($d = 1.08$). 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.

Attitude Moralization and Identity Centrality Across Domains

Figure 2. Figure depicts means and 95% CIs for our 3 key variables across all attitude targets. Moralization and Target IDC are composite items.

Moralization and Identity Centrality

When participants saw their attitude as a reflection of their personal values, did they also see that attitude as part of their identity? 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 expect to positively predict A) targets' identity centrality, B) attitudes' identity centrality, and C) implicit identification with targets.

Target identity centrality. Consistent with our predictions, the moralized attitude composite significantly predicted target identity centrality ($b = 0.15$, $p = < 0.001$, 95% CI [0.13, 0.17], $N_{participants} = 13,548$). The same pattern held for each individual item in the composite ($bs = .33, .18, .13, .08$, $ps < .001$, $N_{participants}$ between 3,326 and 3,420). See Table 2. Figure 3 illustrates these effects. This pattern also holds in separate models that control for attitude importance (bs for moralization = 0.09, 0.08, 0.19, 0.05, 0.11 $ps \leq .010$, $N_{participants}$ between 2,178 and 8,902), certainty (bs for moralization = 0.14, 0.17, 0.32, 0.09, 0.11 $ps < .001$, $N_{participants}$ between 2,190 and 9,002), and extremity (bs for moralization = 0.13, 0.15, 0.28, 0.07, 0.11, $ps < .001$, $N_{participants}$ between 6596 and 26,860).⁹

⁹ Controlling for all three of these variables simultaneously would reduce our sample size by about 93% (from over 13,000 to about 250 participants, still spread across 95 attitude domains), so we instead controlled for each in turn.

Table 2 Predicting Target Identity Centrality from Moralized Attitude Items (See Figure 3)

Outcome: “How much is [A/B] part of your self-concept?”	Moralized Attitude Composite		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	
<u>Fixed Effects</u>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>
Moralized Attitudes	0.15*** (0.01)	<0.001	0.18*** (0.02)	<0.001	0.08*** (0.02)	<0.001	0.13*** (0.01)	<0.001	0.33*** (0.02)	<0.001
Constant	2.55*** (0.08)	<0.001	2.55*** (0.09)	<0.001	2.78*** (0.10)	<0.001	2.66*** (0.09)	<0.001	1.71*** (0.09)	<0.001
<u>Random Effects</u>	σ^2 (se)		σ^2 (se)		σ^2 (se)		σ^2 (se)		σ^2 (se)	
σ^2 (Intercept across participants)	0.13	(0.03)	0.17	(0.05)	0.12	(0.06)	0.06	(0.05)	0.06	(0.06)
σ^2 (Intercept across att. pairs)	0.60	(0.09)	0.61	(0.10)	0.53	(0.11)	0.59	(0.10)	0.48	(0.10)
σ^2 (Slope across participants)	0.01	(0.001)	0.01	(0.002)	0.01	(0.002)	0.01	(0.002)	0.01	(0.002)
σ^2 (Slope across att. pairs)	0.01	(0.001)	0.01	(0.003)	0.02	(0.004)	0.01	(0.002)	0.02	(0.004)
Log-likelihood	-48397		-12338		-12500		-12362		-11937	
Wald χ^2 (degrees freedom)	163.48*** (1)		120.59*** (1)		18.04*** (1)		83.38*** (1)		301.90 *** (1)	
<i>N</i> Level-1 units (responses)	26,967		6,758		6,807		6,776		6,626	
<i>N</i> Level-2 units (participants)	13,548		3,397		3,420		3,405		3,326	
<i>N</i> Level-3 units (att. pairs)	95		95		95		95		95	

Note. Entries derive from 5 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.)

Figure 3. 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 five 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 2.

Attitude identity centrality. The moralized attitude composite also significantly predicted the attitude identity centrality composite ($b = 0.29, p < 0.001, 95\% \text{ CI } [0.26, 0.31], N_{\text{participants}} = 31,119$), suggesting that moralized attitudes tended to be more identity-central than non-moralized attitudes regardless of whether they were positive or negative. To confirm that this trend held for the composites' constituent items, we estimated mixed models in which we regressed each attitude identity centrality item on the moralized attitude item with which it corresponded most closely.¹⁰ Results were consistent with predictions ($bs = 0.36, 0.43, 0.31, 0.52, ps < 0.001, N_{\text{participants}}$ between 1,930 and 1,972). See Table 3 and Figure 4. This pattern also holds in separate models that control for attitude importance (bs for moralization = 0.25, 0.34, 0.40, 0.30, 0.39, $ps \leq .001, N_{\text{participants}}$ between 829 and 13,263), certainty (bs for moralization = 0.30, 0.37, 0.44, 0.31, 0.55, $ps < .001, N_{\text{participants}}$ between 806 and 13,442), and extremity (bs for moralization = 0.26, 0.35, 0.43, 0.31, 0.46, $ps < .001, N_{\text{participants}}$ between 1,929 and 31,066).

Table 3. Predicting Attitude Identity Centrality from Moralized Attitude Items (See Figure 4)

Outcome: Being [accepting/ rejecting] of [A/B] is [consistent/ inconsistent] with my self-concept.	Moralized Attitude Composite		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	
	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>	<i>b</i> (S.E.)	<i>p</i>
Fixed Effects										
Moralized Attitudes	0.29*** (0.01)	<0.001	0.36*** (0.02)	<0.001	0.43*** (0.02)	<0.001	0.31*** (0.02)	<0.001	0.50*** (0.02)	<0.001
Constant	2.95*** (0.07)	<0.001	2.40*** (0.07)	<0.001	1.45*** (0.05)	<0.001	1.13*** (0.05)	<0.001	1.25*** (0.10)	<0.001
Random Effects										
	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across participants)	1.44	(0.03)	0.99	(0.08)	0.57	(0.05)	0.25	(0.04)	0.54	(0.08)
σ^2 (Intercept across attitude pairs)	0.32	(0.06)	0.11	(0.03)	0.01	(0.02)	0.03	(0.02)	0.41	(0.10)
σ^2 (Slope across participants)	0.01	0.001	0.02	(0.004)	0.04	(0.007)	0.06	(0.005)	0.03	(0.003)
σ^2 (Slope across att. pairs)	0.01	0.002	<0.001	<0.001	0.01	(0.004)	0.003	(0.002)	0.01	(0.004)
Log-likelihood	-115553		-7350		-6818		-6229		-6803	
Wald χ^2 (degrees freedom)	460.50*** (1)		463.58*** (1)		364.19*** (1)		378.31*** (1)		572.54*** (1)	
<i>N</i> Level-1 units (responses)	61,955		3,928		3,910		3,842		3,896	
<i>N</i> Level-2 units (participants)	31,119		1,972		1,966		1,930		1,958	
<i>N</i> Level-3 units (attitude pairs)	95		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, 0.39]).
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 ($b = 0.43$, $p < 0.001$, 95% CI [0.39, 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 consistent with my self-concept ($b = 0.31$, $p < 0.001$, 95% CI [0.28, 0.35]).
4. Because of my personal values, I believe that making positive judgments about A/B is acceptable – Being accepting of A/B is consistent with my self-concept ($b = 0.50$, $p < 0.001$, 95% CI [0.46, 0.54]).

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

Figure 4. Lines indicate the predicted values of attitude identity centrality across the full range each moralization predictor. Predictions derive from the fixed portion of five 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 3 includes details on which predictors were used for each attitude IDC outcome variable.

Implicit identification. 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 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 positive effect of moralized attitudes on implicit identification ($b = 0.06, p < 0.001, 95\% \text{ CI } [0.05, 0.07], N_{\text{participants}} = 7,026$). As Figure 5 shows, we obtained the same substantive result (with only minor differences in magnitude) regardless of whether we use the difference between composites or difference scores derived from the composites' constituent items. In sum, the more strongly participants' attitudes were tied to their values, the more strongly those attitudes were tied to the self. With one exception¹¹, this pattern also holds in separate models that control for attitude importance (bs for moralization = 0.05, 0.04, 0.04, 0.04, 0.08, $ps < .006, N_{\text{participants}}$ between 721 and 2,964), certainty (bs for moralization = 0.05, 0.05, 0.05, 0.02 ($n.s. p = .106$), 0.07, other $ps < .001, N_{\text{participants}}$ between 750 and 3,057), and extremity (bs for moralization = 0.05, 0.04, 0.04, 0.04, 0.08, $ps < .001, N_{\text{participants}}$ between 1,721 and 6,993).

¹¹ The difference score for moralized attitude Item 3 (Because of my personal values, I believe that making negative judgments about X/Y is acceptable) did not significantly predict implicit identification with Y over X in the model controlling for attitude certainty ($b = 0.02, p = .106$).

Table 4. Predicting Implicit Identification with A over B from Moralized Attitudes toward A vs. B (See Figure 5)

Outcome: Implicit identification with A over B	Moralized Attitude Composite		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	
	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>
Fixed Effects										
Moralized Attitudes difference score	0.06*** (0.00)	<0.001	0.05*** (0.01)	<0.001	0.05*** (0.01)	<0.001	0.05*** (0.01)	<0.001	0.09*** (0.01)	<0.001
Constant	0.16*** (0.02)	<0.001	0.16*** (0.02)	<0.001	0.17*** (0.03)	<0.001	0.16*** (0.03)	<0.001	0.15*** (0.02)	<0.001
Random Effects										
	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across attitude pairs)	0.05	(0.01)	0.04	(0.01)	0.06	(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	(0.01)	0.00	(0.01)	0.00	(0.00)
Log-likelihood	-4282		-1132		-1124		-1107		-1093	
Wald χ^2 (degrees freedom)	146.53*** (1)		27.86*** (1)		34.67*** (1)		41.95*** (1)		118.02*** (1)	
<i>N</i> Level-1 units (participants)	7026		1732		1766		1737		1791	
<i>N</i> Level-2 units (attitude pairs)	95		95		95		95		95	

Note. Entries derive from 5 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.)

Figure 5. Lines indicate the predicted values of implicit identification across the full range of each of the five moralization variables. These predictions derive from the fixed portion of five 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.

Identity Centrality and Self-Esteem

We next tested our Identity Rubric hypothesis: that self-esteem would be relatively low among participants whose gut feelings contradict their identity-central attitudes. 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.

Target identity centrality. 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 ($b_A = 0.019, p = 0.029$, 95% CI [0.002, 0.036], $N_{\text{participants}} = 624$; $b_B = 0.029, p = 0.002$, 95% CI [0.011, 0.047], $N_{\text{participants}} = 625$). See Table 5 and Figure 6. These results suggest that participants evaluated themselves more negatively when they had negative gut feelings about an identity-central attitude target.

Table 5. Predicting Self-Esteem from Gut Feelings about Identity Central Attitude Targets

Outcome: Self-Esteem	Attitude Target A		Attitude Target B	
	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>
<u>Fixed Effects</u>				
Target Identity Centrality	-0.16* (0.07)	0.020	-0.21** (0.07)	0.001
Gut Feelings about Target	-0.02 (0.03)	0.580	-0.05 (0.03)	0.136
Target IDC x Gut Feelings	0.02* (0.01)	0.029	0.03** (0.01)	0.002
Constant	4.85*** (0.21)	<0.001	5.04*** (0.19)	<0.001
<u>Random Effects</u>				
	σ^2	(se)	σ^2	(se)
σ^2 (Intercept across attitude pairs)	0.01	(-)	<0.01	(-)
σ^2 (Random slope for Target IDC)	<0.01	(-)	<0.01	(-)
σ^2 (Random slope for Gut Feelings)	<0.01	(-)	<0.01	(-)
σ^2 (Random slope for T IDC x GF)	<0.01	(-)	<0.01	(-)
Log-likelihood	-844.93		-844.17	
Wald χ^2 (degrees freedom)	10.1* (3)		14.45** (3)	
<i>N</i> Level-1 units (participants)	624		625	
<i>N</i> Level-2 units (attitude pairs)	95		95	

Estimates derive from the two multi-level linear models depicted in Figure 6, each of which included random slopes and intercepts that varied across attitude stimuli. Stata 16 was unable to estimate standard errors for the variance in slopes and intercepts across stimuli, most likely because of the small number of observations per group (about 7) or because of the near-0 estimated variance in parameters across stimuli.

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

Gut Feelings, Target IDC, and Self-Esteem

Figure 6. 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 stimuli.

Attitude identity centrality. We next examined attitude IDC—the extent to which participants viewed accepting or rejecting A or B as consistent or inconsistent with their self-concept. We did not find significant evidence for an interaction between gut feelings and attitude IDC ($b_A = 0.007$, $p = 0.142$, 95% CI [-0.003, 0.017], $N_{\text{participants}} \equiv 1,463$; $b_B = 0.003$, $p = 0.521$, 95% CI [-0.007, 0.013], $N_{\text{participants}} \equiv 1,468$). See Table 6 and Figure 7.

Table 6. Predicting Self-Esteem from Gut Feelings and Identity Central Attitudes

Outcome: Self-Esteem	Attitude Target A		Attitude Target B	
	<i>b</i> (se)	<i>p</i>	<i>b</i> (se)	<i>p</i>
<u>Fixed Effects</u>				
Attitude Identity Centrality	-0.04 (0.04)	0.284	-0.01 (0.03)	0.811
Gut Feelings about Target	-0.00 (0.02)	0.833	0.02 (0.02)	0.381
Attitude IDC x Gut Feelings	0.01 (0.01)	0.142	0.00 (0.00)	0.521
Constant	4.66*** (0.15)	<0.001	4.52*** (0.13)	<0.001
<u>Random Effects</u>				
	σ^2 (se)		σ^2 (se)	
σ^2 (Intercept across attitude pairs)	0.01 (-)		<0.01 (-)	
σ^2 (Random slope for Attitude IDC)	<0.01 (-)		<0.01 (-)	
σ^2 (Random slope for Gut Feelings)	<0.01 (-)		<0.01 (-)	
σ^2 (Random slope for A IDC x GF)	<0.01 (-)		<0.01 (-)	
Log-likelihood	-1966.7		-1974.9	
Wald χ^2 (degrees freedom)	10.24* (3)		14.01** (3)	
<i>N</i> Level-1 units (participants)	1463		1468	
<i>N</i> Level-2 units (attitude pairs)	95		95	

Estimates derive from the two multi-level linear models depicted in Figure 7, each of which included random slopes and intercepts that varied across attitude stimuli. Stata 16 was unable to estimate standard errors for the variance in slopes and intercepts across stimuli, most likely because of the small number of observations per group (about 15) or because of the near-0 estimated variance in parameters across stimuli.

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

Gut Feelings, Attitude IDC, and Self-Esteem

Figure 7. 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.

Implicit identification. 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 ($b = 0.009$, $p = 0.767$, 95% CI [-0.046, 0.062], $N_{\text{participants}} = 346$; $N_{\text{domains}} = 95$) and is depicted in Figure 8.

Table 7. Predicting Self-Esteem from Gut Preferences and Implicit Identification with Attitude Targets

Outcome: Self-Esteem		
	<i>b</i> (se)	<i>p</i>
Implicit Identification with A over B	-0.07 (0.12)	0.566
Gut Preference for A over B	0.01 (0.02)	0.398
Implicit ID x Gut Preference	0.01 (0.03)	0.767
Constant	4.64*** (0.06)	<0.001
<u>Random Effects</u>		
	σ^2	(se)
σ^2 (Intercept across attitude pairs)	<0.01	(-)
σ^2 (Random slope for Implicit ID)	0.18	(-)
σ^2 (Random slope for Gut Preference)	<0.01	(-)
σ^2 (Random slope for I ID x Gut Pref)	<0.01	(-)
Log-likelihood	-486.85	
Wald χ^2 (degrees freedom)	0.59 (3)	
<i>N</i> Level-1 units (participants)	346	
<i>N</i> Level-2 units (attitude pairs)	95	

Estimates derive from the multi-level linear model depicted in Figure 8, which included random slopes and intercepts that varied across attitude pairs. Stata 16 was unable to estimate standard errors for the variance in slopes and intercepts across stimuli, most likely because of the small number of observations per group (3.6) or because of the near-0 estimated variance in parameters across stimuli. (†*p* < 0.10. **p* < 0.05. ***p* < 0.01. ****p* < 0.001.)

Figure 8. 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.

Proposed 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. 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). In our Stage 2 submission, we will also investigate how people evaluate themselves when their spontaneous affective reactions contradict attitudes that are important, extreme, certain, or moralized.

First, we will examine each of these strength-related attitude attributes separately. Second, we will individually compare each attribute with attitude identity centrality. Third, we

will individually compare each attribute with target identity centrality. See Table 8 for a summary of these models and the predictors they will include. These analyses will clarify whether the effects we predict for identity centrality are unique and specific to identity centrality *per se*.

We have not yet performed these tests of alternative rubrics because power analysis (reported in the discussion) suggest that our exploratory tests of the Identity Rubric hypothesis are underpowered, and adding covariates would exacerbate this problem.

Table 8. Proposed Tests of Alternative “Rubrics” for Self-Esteem

Rubric: Moralized Attitudes	Rubric: Important Attitudes	Rubric: Certain Attitudes	Rubric: Extreme Attitudes
<u>Predictors in First Model</u>	<u>Predictors in First Model</u>	<u>Predictors in First Model</u>	<u>Predictors in First Model</u>
Moralized Attitudes	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
<u>Predictors in Second Model</u>	<u>Predictors in Second Model</u>	<u>Predictors in Second Model</u>	<u>Predictors in Second Model</u>
[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
<u>Predictors in Third Model</u>	<u>Predictors in Third Model</u>	<u>Predictors in Third Model</u>	<u>Predictors in Third Model</u>
[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 would use to investigate how people evaluate themselves when their spontaneous affective reactions contradict attitudes that are important, extreme, certain, or moralized (vs. identity-central). The outcome in every model will be participants’ self-esteem. All models will allow slopes and intercepts to vary randomly across attitude stimuli (unless models fail to converge, in which case we will only allow the intercept to vary).

“Actual feelings” were omitted from models including the moralized attitude measure because the moralized attitude measure these three variables were not confounded with valence, unlike the AIID measures of moralized attitudes and identity centrality, importance “actual feelings” are effectively “baked in” to the AIID measures of moralized attitudes and identity centrality

Discussion of Exploratory Results

Summary

Evidence from the exploratory subset of the AIID dataset is consistent with the Morality-Identity hypothesis: moralized attitudes were consistently more central to participants' identities than non-moralized attitudes, and participants even rated the targets of these attitudes (e.g., McDonald's *per se*) as more central to their identities when they moralized favorable reactions to them. We find weaker and inconsistent evidence for the Identity Rubric hypothesis. Self-esteem was sometimes, but not robustly, related to the degree of consistency between participants' gut feelings and their identity-central attitudes. However, these analyses may have been underpowered. We plan to use the confirmatory subset of the AIID data to resolve this complicated pattern of results.

Limitations

The planned missingness design of the AIID dataset introduces some important measurement challenges. Both attitude moralization and attitude identity centrality were measured with a single item for all participants. Single items are known to be less reliable than multi-item scales (Eagly & Chaiken, 1993). The large samples counter some of the imprecision caused by low reliability. However, the single item that was available differed across participants, raising the possibility that substantive differences in the content across these items produced noise or bias in our estimates. The same problem precludes analyses that would help us to assess the internal reliability or dimensionality of the corresponding composite variables. For now, we test our hypotheses separately for each item and to avoid the assumption that either the attitude moralization or attitude IDC items tap a single, one-dimensional construct. Our proposed validation study, however, will reveal whether these disparate measures can be reasonably

combined and how cleanly they assess moralization and identification. Thus, validation data will reveal which of these disparate tests provide the most appropriate tests of our hypotheses.

Another potential issue stems from the diversity of attitude targets included in the AIID dataset. This diversity mostly works to our advantage; the fact that the dataset includes responses to both trivial, everyday objects (e.g., boxers, briefs) and to abstract, generally desirable values (e.g., giving, wisdom) lowers the floor and raises the ceiling between which we can observe variance in moralization and identity centrality. However, some differences across stimuli may change the way participants interpret and answer key questions, introducing noise and measurement error.

Planned Confirmatory Analyses: Moralization, Identity Centrality, and Self-Esteem

The analyses reported above rely on the exploratory subset of the AIID data—about 15% of the total sample. Once we have access to the remaining 85% of the AIID dataset, we propose to repeat all of our exploratory analyses—exactly as performed above—to determine whether they replicate in the larger dataset.

Conclusion

Significance of Proposed Analyses

Our goals with the present study are three-fold. First, we aim to understand the role of moralized attitudes 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 test this possibility.

Second, we aim to clarify the nature and function of moralized attitudes. Moralized attitudes have a number of unique characteristics. They are highly resistant to persuasion, personally important, and evocative of strong emotion (Skitka & Wisneski, 2011). They motivate political participation (Morgan, Skitka, & Wisneski, 2010; Skitka & Bauman, 2008) and intolerance of disagreement (Skitka & Bauman, 2008), and can even undermine the perceived legitimacy of political institutions (Skitka & Mullen, 2002). Here we explore another potentially unique role of moralized attitudes—as a core component of the self and as a basis for self-evaluation. Indeed, moralized attitudes’ identity centrality might explain many of their other special properties. Future work might examine identity centrality as a mediator of other known effects of moralization.

Finally, we aim to reveal (or rule out) potential parallels between how people evaluate themselves and how they evaluate others. If moralized attitudes follow the patterns that we predict, our results would suggest that a common set of considerations—moral beliefs and values—inform people’s evaluations of themselves and others. This would be a valuable insight. It would be perfectly reasonable to expect different psychological mechanisms to govern evaluations of the self and others, but the more parsimonious possibility that people evaluate human agents *in general* on the basis of their morality is worth exploring.

Interpreting Unanticipated Outcomes

The confirmatory analyses that we have proposed will further our aims regardless of whether the results are consistent with our hypotheses. Moralized attitudes may *not* play a special role in self-evaluation. Their identity centrality may not matter. The size and scope of the AIID dataset—which we have only begun to explore here—would put us in a uniquely powerful position to draw substantive conclusions from these surprising null results.

Our exploratory results already provide robust evidence consistent with our Morality-Identity hypothesis. Simulations suggest that our tests of this hypothesis had over 99% power to detect the smallest effects of theoretical interest to us in this case ($b = .10$),¹² so we do not expect confirmatory analyses to overturn these findings. Nevertheless, null results in the much larger confirmatory subset of the AIID data would call into serious question the identity centrality of moralized attitudes and would suggest that identity processes could *not* account for moralized attitudes' unique properties. If we do obtain null results (i.e., a non-significant relation between attitude moralization and identity centrality), we will conduct a sensitivity analysis to determine whether this “null” result is indeed significantly smaller than our smallest effect size of theoretical interest. If the 95% confidence interval for our focal coefficient b does not include 0.1, we would interpret our results as consistent with the null hypothesis. This coefficient corresponds to a half-point increase in identity centrality across the full range of the moralized attitude composite.

In contrast, simulations suggest that our exploratory tests of the Identity Rubric hypothesis were not well-powered. We had 49% power to detect small interaction effects between gut feelings and target IDC and about 33% power to detect small interactions between gut feelings and the decomposed attitude IDC items (by “small interactions”, we mean interaction coefficients that yield a small simple slope ($b = 0.10$) for gut feelings when identity centrality is at its maximum). Power was lower still (19%) to detect the equivalent interaction between the gut feeling difference score and implicit identification. Analyses using the larger confirmatory subset of the data will resolve these power issues, affording 90-99% power to detect small interactions between gut feelings and both explicit IDC measures and 83% power to

¹² These simulations are described in detail in the supplemental file uploaded with our manuscript, “Power Analyses R1,” also available at https://osf.io/6ckns/?view_only=794bdcfca57f40bb9ecfca0389c6debc.

detect an equivalent interaction between gut feelings and implicit identification.¹³ With this considerable increase in power, null effects with respect to our Identity Rubric hypothesis would be estimated with a high degree of precision. If our focal interaction does not significantly differ from 0, or if the coefficient falls below our threshold for a theoretically interesting effect size (i.e., $b_{\text{interaction}} = 0.02$ for models including attitude or target IDC, $b_{\text{interaction}} = 0.0359$ for models including implicit identification), we will conduct a sensitivity analysis. If the 95% CI for the interaction falls short of the minimum threshold, we would interpret our results as consistent with the null hypothesis. This null pattern would be interesting in its own right, suggesting that self-esteem is 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.

The AIID is the largest dataset of its kind assembled to date. Even restricting our analyses to those participants who completed the Rosenberg Self-Esteem scale will permit us a total sample size of over 3,000 individuals, including responses to dozens of attitude domains. There is no better source of data that we could use to examine the role of moralized attitudes in individuals' identities and their evaluations of themselves. Our exploratory analyses already provide strong evidence that individuals' moralized attitudes are an important part of how they see themselves. But do they also inform individuals' judgments of their own self-worth? And if so, do they do so *because* of their centrality to the self-concept? In the confirmatory AIID data, we will examine these questions and learn more of peoples' capacity to deny, rationalize, or come to terms with the ways in which they fall short of the people they aspire to be.

References

¹³ For a detailed description of these simulations, see the supplemental file uploaded with our manuscript, "Power Analyses R1," also available at https://osf.io/6ckns/?view_only=794bdcfca57f40bb9ecfca0389c6debc.

- Aquino, K., & Reed, A. (2002). The Self-Importance of Moral Identity. *Journal of Personality and Social Psychology*, 83, 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](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 Petty, R. & Krosnick, J. A. (Eds.) *Attitude strength: Antecedents and consequences*, 159-189.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11, 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). Orlando, FL: Academic Press.
- Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition*, 25, 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, 148-168.
<https://dx.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, 373-377.
<https://doi.org/10.1177%2F0963721412456842>
- Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.
- Higgins, E. T. (1989). Self-discrepancy theory: What patterns of self-beliefs cause people to suffer?. In *Advances in experimental social psychology* (Vol. 22, pp. 93-136). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60306-8](https://doi.org/10.1016/S0065-2601(08)60306-8)
- Higgins, E. T., Roney, C. J., Crowe, E., & Hymes, C. (1994). Ideal versus ought predilections for approach and avoidance distinct self-regulatory systems. *Journal of personality and social psychology*, 66, 276-286. <https://doi.org/10.1037/0022-3514.66.2.276>
- Janoff-Bulman, R., Sheikh, S., & Baldacci, K. G. (2008). Mapping moral motives: Approach, avoidance, and political orientation. *Journal of Experimental Social Psychology*, 44, 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, 701-713. <https://doi.org/10.1177%2F0146167211400208>
- 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, 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, 660-674. <https://doi.org/10.1037/pspa0000050>
- Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110, 675-709. <https://doi.org/10.1037/pspa0000046>
- Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. *Attitude strength: Antecedents and consequences*, 1, 1-24.
- 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, 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, 302-318.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390. <https://doi.org/10.1037/0033-2909.114.2.376>
- McFerran, B., Aquino, K., & Duffy, M. (2010). How personality and moral identity relate to individuals' ethical ideology. *Business Ethics Quarterly*, 20, 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, 344-357. <https://doi.org/10.1111/j.1751-9004.2010.00263.x>

- Morgan, G. S., Skitka, L. J., & Wisneski, D. C. (2010). Moral and religious convictions and intentions to vote in the 2008 presidential election. *Analyses of Social Issues and Public Policy*, 10, 307-320. <https://doi.org/10.1111/j.1530-2415.2010.01204.x>
- Muthén, L.K. and Muthén, B.O. (2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén
- Prentice, M., Jayawickreme, E., Hawkins, A., Hartley, A., Furr, R. M., & Fleenor, W. (2019). Morality as a Basic Psychological Need. *Social Psychological and Personality Science*, 10, 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, 386-396. <https://doi.org/10.1016/j.jesp.2006.12.008>
- Reed, I. I., & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of personality and social psychology*, 84, 1270-1286. <https://doi.org/10.1037/0022-3514.84.6.1270>
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53, 550-562. <https://doi.org/10.1037/0022-3514.53.3.550>
- Simonsohn, Uri and Simmons, Joseph P. and Nelson, Leif D., Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications (November 24, 2015). Available at SSRN: <https://ssrn.com/abstract=2694998> or <http://dx.doi.org/10.2139/ssrn.2694998>

- Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, 4, 267-281. <https://doi.org/10.1111/j.1751-9004.2010.00254.x>
- Skitka, L. J., & Bauman, C. W. (2008). Moral conviction and political engagement. *Political Psychology*, 29, 29-54. <https://doi.org/10.1111/j.1467-9221.2007.00611.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, 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.), *Advances in Political Psychology*, 35, 95 – 110.
- Skitka, L. J., & Mullen, E. (2002). Understanding judgments of fairness in a real-world political context: A test of the value protection model of justice reasoning. *Personality and Social Psychology Bulletin*, 28, 1419-1429. <https://doi.org/10.1177%2F014616702236873>
- Skitka, L. J. & Wisneski, D. C. (2011). Moral conviction and emotion. *Emotion Review*, 3, 328 - 330. <https://doi.org/10.1177%2F1754073911402374>
- 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, 884-895. <https://doi.org/10.1037/xge0000317>
- 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.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25, 1115-1125.
<https://doi.org/10.1177%2F01461672992512005>
- Winterich, K. P., Mittal, V., & Ross Jr, W. T. (2009). Donation behavior toward in-groups and out-groups: The role of gender and moral identity. *Journal of Consumer Research*, 36, 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, 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, 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, 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, 1251-1263.
<https://doi.org/10.1177%2F01461672982412001>
- Wojciszke, B., & Sobiczewska, P. (2013). Memory and self-esteem: The role of agentic and communal content. *Social Psychology*, 44, 95-103. <https://doi.org/10.1027/1864-9335/a000149>

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, 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

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, Krosnick, Berent, and Fabrigar, 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