Journal of Experimental Psychology: General
2018, Vol. 147, No. 11, 1748–1761

0096-3445/18/$12.00 http://dx.doi.org/10.1037/xge0000404

An Empirically-Derived Taxonomy of Moral Concepts

Justin F. Landy and Daniel M. Bartels
University of Chicago

We propose that methods from the study of category-based induction can be used to test the descriptive accuracy of theories of moral judgment. We had participants rate the likelihood that a person would engage in a variety of actions, given information about a previous behavior. From these likelihood ratings, we extracted a hierarchical, taxonomic model of how moral violations relate to each other (Study 1). We then tested the descriptive adequacy of this model against an alternative model inspired by Moral Foundations Theory, using classic tasks from induction research (Studies 2a and 2b), and using a measure of confirmation, which accounts for the baseline frequency of these violations (Study 3). Lastly, we conducted focused tests of combinations of violations where the models make differing predictions (Study 4). This research provides new insight into how people represent moral concepts, connecting classic methods from cognitive science with contemporary themes in moral psychology.

Keywords: moral judgment, inductive reasoning, concepts and categories, taxonomic models

Supplemental materials: http://dx.doi.org/10.1037/xge0000404.supp

The psychology of moral judgment has been a very active area of research in recent years (see, e.g., Bartels, Bauman, Cushman, Pizarro, & McGraw, 2016; Sinnott-Armstrong, 2008), with several theories proposing a handful of discrete types of moral violations (e.g., Janoff-Bulman & Carnes, 2013; Graham, Haidt, & Nosek, 2009; Rozin, Lowery, Imada, & Haidt, 1999). However, this research has generally not examined whether these theories accord with laypeople's understandings of moral concepts. In this article, we use methods from the study of category-based induction (e.g., Feeney & Heit, 2007; Osherson, Smith, Wilkie, López, & Shafir, 1990) to investigate this question. We use inductive judgments of the likelihood of different behaviors, given prior behaviors, to derive a model of how moral violations relate to each other and compare this model to another that is inspired by Moral Foundations Theory (MFT; Graham et al., 2009; Haidt, 2012; Haidt & Joseph, 2004; Iyer, Koleva, Graham, Ditto, & Haidt, 2012). We consider MFT to be a useful case study in the use of inductive judgments to study representations of moral concepts, because it has inspired a great deal of research on a diverse array of topics (e.g., persuasion, Day, Fiske, Downing, & Trail, 2014; personality disorders, Glenn, Iyer, Graham, Koleva, & Haidt, 2009; life narratives, McAdams et al., 2008; victim-blaming, Niemi & Young, 2016; economic behavior, Schier, Ockenfels, & Hofmann, 2016), and it posits discrete categories of moral violations that can be studied using these methods.

MFT argues that we are attuned to certain patterns of behavior that prompt intuitive judgments of approval or disapproval (Haidt & Joseph, 2004). These intuitions are elaborated into clusters of virtues that are observed cross-culturally. The earliest version of MFT (Haidt & Joseph, 2004) considered four of these "moral foundations": harm prevention ("care"), fairness, respect for and obedience to authority, and bodily and spiritual purity. Later work added a fifth foundation—loyalty to important in-groups, and drew a social-functional distinction between care and fairness—the "individualizing" foundations—and loyalty, authority, and purity—the "binding" foundations (Graham et al., 2009; Haidt & Graham, 2007). These different clusters of foundations are thought to represent different strategies for suppressing selfishness to allow for group living (Haidt & Kesebir, 2010). The individualizing foundations regulate behavior by instilling respect for others' rights and welfare, whereas the binding foundations do so by limiting freedoms and prescribing roles for group members to fulfill (Haidt, 2008). Recent work has proposed a sixth moral foundation, liberty, which has not yet received the same level of empirical scrutiny as the five that preceded it (Iyer et al., 2012, see also Graham et al., 2011). The liberty foundation fits within the superordinate category of individualizing foundations, which are about the rights of individuals, rather than the good of collectives (J. Graham, personal communication, October 5, 2017; see also Haidt, 2012).

This article was published Online First March 19, 2018.

Justin F. Landy, Center for Decision Research, University of Chicago; Daniel M. Bartels, Department of Marketing, University of Chicago.

We thank Halley Bayer and Johannes de Nova for their assistance in conducting this research. We also thank Stephanie Chen, Morteza Dehghani, Geoffrey Goodwin, Jesse Graham, Craig Joseph, Peter McGraw, Doug Medin, and his lab group; Daniel Osherson, Lance Rips, and his lab group; Nina Strohminger, and the members of the Morality Research Lab for their helpful feedback on this project. We gratefully acknowledge financial support from the Neubauer Family Faculty Fellows Program and the Kilts Center for Marketing at the University of Chicago Booth School of Business. Portions of this research were previously presented at the Annual Meeting of the Society for Judgment and Decision Making, Boston, MA, November, 2016, the Annual Meeting of the Society for Philosophy and Psychology, Austin, TX, June, 2016, the International Conference on Thinking, Providence, RI, August, 2016, and the Annual meeting of the Cognitive Science Society, Philadelphia, PA, August, 2016.

Correspondence concerning this article should be addressed to Justin F. Landy, Center for Decision Research, University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637. E-mail: justinlandy@chicagobooth.edu

These multiple virtues give rise to distinct categories of violations, with qualitatively different kinds of actions violating care versus purity, for instance (Chakroff, Dungan, & Young, 2013; Dungan, Chakroff, & Young, 2017; Young & Saxe, 2011).

It is plausible that the social-functional distinction between individualizing and binding moral foundations is also a psychological distinction that people make. One way to represent this would be a hierarchical knowledge structure, as presented in Figure 1. In this representation, the individualizing foundations belong to one superordinate category of virtues, whereas the binding foundations belong to a separate superordinate category (we present the liberty foundation, and its conceptual link to its superordinate category, individualizing foundations, as dotted lines, to represent the less established status of this foundation within MFT). However, MFT is usually presented as a theory of evolutionary and cultural psychology, not of concepts and categories, so it remains an open question whether or not people's mental representations of moral concepts distinguish them along this social-functional divide.

In what follows, we use people's inductive judgments about the likelihood of different behaviors to derive a model of mental representations of moral concepts. In doing so, we demonstrate how methods from the study of induction can be used to test the descriptive accuracy of models of moral judgment.

Category-Based Induction and Morality

The study of category-based induction often assumes that concepts are organized taxonomically, and we accept this as a working assumption here.[1] On this assumption, the strength of inductive inferences that a person makes from one object to another depends on how closely related the objects are in that person's taxonomic representation. Consider an example, adapted from Osherson et al. (1990). Given the premise "robins use serotonin as a neurotransmitter", the conclusion "sparrows use serotonin as a neurotransmitter" is usually considered more likely to be true than "geese use serotonin as a neurotransmitter." This is because robins and sparrows are closer to one another in people's taxonomies of birds than are robins and geese. Robins and sparrows might belong to the superordinate category "songbirds," whereas geese would belong to a separate superordinate category, perhaps "waterfowl." Robins and geese unite only at a higher level of the taxonomy (presumably the top-level category, "birds"), and therefore inductive inferences from robins to geese are not especially strong.

We applied this same logic to taxonomies of moral violations. Consider the premise "Joe committed a violation of fairness" (henceforth, a "fairness violation", etc.). If an MFT-like taxonomy as shown in Figure 1 is a viable model of people's representations of moral concepts, then the conclusion "Joe would commit a care violation" should be considered more likely than the conclusion "Joe would commit an authority violation," because fairness and care belong to the same superordinate category (i.e., individualizing foundations), whereas authority belongs to a separate superordinate category (i.e., binding foundations) and only unites with fairness at the top-level category, moral virtues.[2]

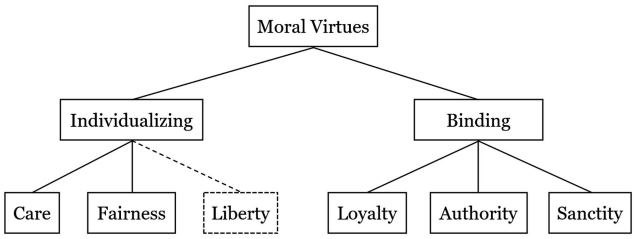

Figure 1. A plausible representation of moral concepts based on Moral Foundations Theory.

We examined the structure of people's representations of moral concepts across five studies. In Study 1, participants rated the likelihood that a person would engage in a wide variety of actions that violate the six moral foundations (conclusions), given information about previous behavior (premises). We found that there is consensual knowledge that our participants drew upon in making these judgments. From their likelihood ratings, we derived a taxonomy of moral concepts to model this shared knowledge. This taxonomy does not resemble an MFT-like taxonomy, but it has an interpretable structure. In Studies 2a and 2b, participants indicated which premises most strongly supported which conclusions, in tasks adapted from classic research on category-based induction. Next, in Study 3, we modified the task from Study 1 to account for differences in baseline frequency among our stimuli. Finally, in Study 4, we focused on cases where the derived taxonomy and an MFT-like taxonomy make differing predictions about membership in superordinate categories. Across Studies 2 through 4, participants' inductive judgments more closely resembled the predictions of the derived taxonomy than an MFT-like taxonomy.

Study 1

Study 1 had two aims. The first was to assess whether there is enough agreement about the conceptual relatedness of different kinds of moral violations to warrant development of a model of people's taxonomies of moral concepts. If there is, the second aim was to use exploratory analyses to extract the taxonomic structure of representations of moral concepts.

Method

Stimuli. We felt that it was important to use stimuli where the violations of each moral foundation differed primarily in the moral foundation that characterized them, rather than on other factors like their overall moral wrongness. We worried that if the violations of some foundations were more severe than others, then participants' likelihood ratings could potentially just group together by wrongness—our participants might infer that a person who had committed a particularly bad prior act (premise) is just a

[1] Although this project follows other category-based induction research that assumes taxonomic representations (see, e.g., Carey, 1985; Choi, Nisbett, & Smith, 1997; Gelman & Coley, 1990; Heit, 2000; López, 1995; Osherson et al., 1990), there are, of course, good alternative ideas about how concepts are represented (see, e.g., Rips, Smith, & Medin, 2012; Murphy, 2004).

[2] It is not clear that MFT necessarily predicts that people's representations of moral concepts will be organized by social-functional categories. So, we use "MFT" to refer to Moral Foundations Theory as it is typically articulated, and "MFT-like" to describe the taxonomy presented in Figure 1.

morally worse person, prone to especially bad behavior. They might therefore infer that this person would be more likely to commit a particularly bad subsequent act (conclusion) for reasons other than the specific virtue that each act violates. To address this concern, we conducted five norming studies (total $N = 1{,}267$) on 244 behavioral descriptions. The studies were conducted on Amazon Mechanical Turk, the same population that participated in our main studies. Most of the stimuli included in the norming studies were original, or were modified forms of the Moral Foundations Vignettes (Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015), though a minority came from other research. Stimuli and details about the norming studies are presented in the online supplemental materials.

We selected seven behaviors violating each moral foundation for our stimulus set (e.g., a person drives past a clearly injured man on an empty road [care], edges out another person in a long line [fairness], forces their daughter to enroll as a premed student in college [liberty], makes critical comments about their home country [loyalty], sends out an email calling their boss an "idiot" [authority], or views deviant pornography [purity]). These stimuli uniquely exemplify the moral foundations, and provide broad conceptual coverage of each one (e.g., the liberty stimuli include both overbearing parents and overreaching politicians; see documentation in the online supplemental materials). Also, the mean moral wrongness ratings for the six moral foundations are closely equated (5.20 to 5.26, on a 1 to 9 scale in pretests). We also included seven nonmoral actions that extensive pretesting found to be morally inert (e.g., "a person goes parasailing"). Last, we included seven counternormative actions that do not exemplify any moral foundation (e.g., "while in a rush, a person bumps into someone on the street, but does not say 'excuse me'"). These actions largely consist of violations of polite etiquette.

Participants. Four hundred twenty-five participants were recruited online through Amazon Mechanical Turk. Throughout this article, we excluded participants for failing "captcha" verifications (suggesting they were automated "bot" programs), failing to reach the end of a study, or failing attention checks. After exclusions, we retained a final sample of 372 (195 female, $M_{Age} = 37.47$, $SD_{Age} = 11.36$). Because this study is exploratory in its methods, we sought to recruit a fairly large sample of approximately 350 participants. This target sample size was determined before data collection.[3]

Materials and procedure. Each participant made 64 likelihood judgments, one for each possible premise/conclusion combination of the eight conceptual categories in our stimulus set (e.g., authority/authority, authority/nonmoral). We had participants judge the likelihood of other behaviors (rather than, for instance, making similarity judgments) because a central goal in understanding people's moral virtues is to predict their likely intentions and behaviors (see, e.g., Cottrell, Neuberg, & Li, 2007; Landy, Piazza, & Goodwin, 2016; Landy & Uhlmann, in press; Pizarro & Tannenbaum, 2011; Wojciszke, Bazinska, & Jaworski, 1998). So, this seemed like a natural task for examining representations of virtues and violations.

Premises and conclusions were randomly sampled for each question, with the restriction that the premise and conclusion could not be the same action. Questions took the following form: "A person edges out another person in a long line. Given this information, how likely is it that, if they were driving along an empty road and saw a man who was clearly injured, this person would drive past the man and not stop to help him?" This is one of 7 premises × 7 conclusions = 49 possible fairness/care questions. Likelihood ratings were made using a sliding scale (0% = "There is no chance this person would do this"; 100% = "This person would definitely do this").

All studies reported in this article were reviewed and approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board. Data and materials from all studies reported in this article are publicly available on the Open Science Framework at https://osf.io/k5mpr/.

Results

Cultural consensus analysis. The first aim of this study was to examine whether people agree on the relations between the different kinds of actions that participants judged. To test this, we used the informal version of the Cultural Consensus Model (a part of Cultural Consensus Theory [CCT]; Romney, Batchelder, & Weller, 1987; Weller, 2007). CCT recommends statistical techniques to assess, by the degree of consistency across participants' responses, whether they are drawing on the same shared knowledge in making their judgments. If there was not consensus across participants in our data, this would undermine the goal of extracting a taxonomic model, as different participants would be representing these concepts in vastly different ways.

Following the methods of CCT, we transposed our data such that the 64 likelihood judgments that participants made (fairness/care, authority/nonmoral, etc.) were the rows, and participants were the columns. We then used unrotated principal components analysis to examine the degree of consensus. The first and second factors extracted had eigenvalues of 81.09 and 17.00. In CCT, it is generally accepted that a ratio of the first to the second eigenvalue of 3.0 or greater indicates consensus (see Weller, 2007). This ratio in this analysis was 4.77, suggesting that our participants drew on shared, consensual knowledge in making their likelihood judgments. We now turn to characterizing the structure of this consensual representation.

Exploratory analyses. We computed a bidirectional measure of conceptual relatedness between categories of actions by averaging the likelihood judgments made by a participant within each pair of categories. For instance, if a participant rated the likelihood of committing a fairness violation, knowing that a person had committed a care violation, as 70%, and the likelihood of committing a care violation, knowing that a person had committed a fairness violation, as 50%, that participant's fairness/care relatedness score would be (70% + 50%)/2 = 60%.

Mean conceptual relatedness scores are presented in Table 1. Note that within-foundation relatedness scores (presented on the diagonal) are generally larger than between-foundation scores (presented off-diagonal), suggesting that our judgment task captures the conceptual relatedness between different actions. Also,

[3] Participants could not take part in more than one study in this project (including the norming studies described in the online supplemental materials), with one exception. Studies 2a and 2b were run over 1 year after the norming studies and Studies 1, 3, and 4 in response to feedback from colleagues on an earlier version of this article. We therefore did not exclude participants from the earlier studies from Studies 2a and 2b (though each participant could take part in only Study 2a or Study 2b, not both).

Table 1
Mean Conceptual Relatedness Scores, Defined as the Mean Likelihood Judgment Within Category Pairs, Study 1

Category	1	2	3	4	5	6	7	8
1. Care	60%	52%	45%	56%	52%	37%	38%	53%
2. Fairness		63%	46%	57%	55%	35%	42%	53%
3. Liberty			60%	40%	39%	28%	37%	41%
4. Authority				68%	55%	38%	40%	56%
5. Loyalty					61%	35%	39%	45%
6. Purity						36%	32%	36%
7. Nonmoral							54%	41%
8. Counternormative								57%
9. Mean relatedness	49%	50%	42%	51%	48%	35%	40%	48%

notably, liberty violations, purity violations, and nonmoral actions tend to have lower mean conceptual relatedness scores than other categories (mean scores: 42%, 35%, and 40%; other categories ranged from 48% to 51%). This suggests that these three types of actions are somewhat distinct from other types of actions in people's taxonomies of moral concepts.

We next used hierarchical cluster analysis and multidimensional scaling to explore these patterns (see Medin et al., 2006, for application of these methods to folk taxonomies of natural kinds). First, we submitted the mean relatedness scores to a hierarchical cluster analysis using average linkage between groups. Average linkage is considered to be a compromise between single linkage and complete linkage that minimizes the problems that these methods can create (Sokal & Michener, 1958; Yim & Ramdeen, 2015). Consistent with the pattern of means in Table 1, violations of care, authority, fairness, and loyalty, and counternormative actions, were close to one another in Euclidean space and clustered early in the analysis. In contrast, violations of liberty and purity, and nonmoral actions, were quite distant from all other categories.[4] Figure 2 presents a dendrogram illustrating this analysis.

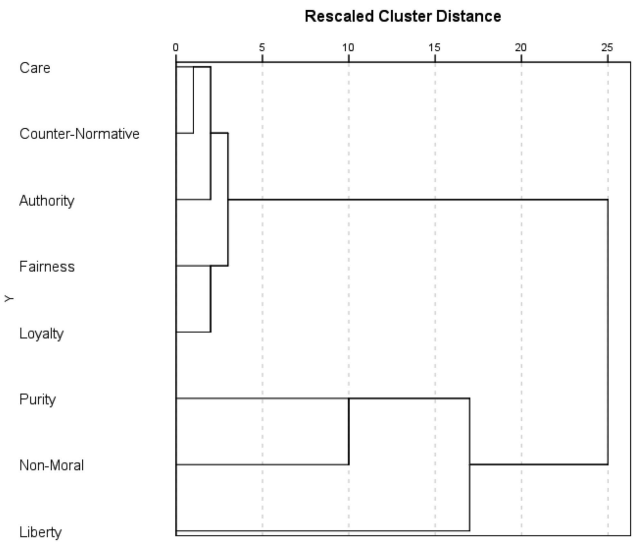

Figure 2. Dendrogram illustrating hierarchical cluster analysis (average linkage) of relatedness scores, defined as the mean likelihood judgment within category pairs. The *x*-axis represents squared Euclidean distances between agglomerated clusters.

We confirmed this result by subtracting relatedness scores from 100%, and submitting the resulting dissimilarity scores to multidimensional scaling, treating the dissimilarity scores as ordinal variables.[5] Figure 3 presents the two-dimensional solution. Violations of care, authority, fairness, and loyalty, and counternormative actions, are quite close to one another in the resultant two-dimensional space, with liberty violations, and especially purity violations and nonmoral actions, more distant. Model stress was .08, which is generally considered acceptable (see, e.g., Kruskal, 1964a, 1964b; Rosenberg, Nelson, & Vivekananthan, 1968), though the model stress for the three-dimensional solution was lower, at .04. A four-dimensional solution did not converge because there were too many parameters to estimate. So, the three-dimensional solution provides the best available model of these data. In this model, and in agreement with the analyses above, liberty violations, purity violations, and nonmoral actions were quite distant from all other categories, including each other (mean Euclidean distances 2.42, 2.98, and 2.49, respectively, all other categories < 2.11).[6]

Discussion

These analyses suggest that care, fairness, authority, and loyalty violations, and impolite actions, are closely related in the representations of moral concepts that our participants shared. Liberty

[4] We conducted a series of robustness checks on this analysis. The results are similar if conceptual relatedness is defined as the minimum, maximum, or product of likelihood judgments within a category pair, or if single or complete linkage is used to join clusters. Also, if liberty violations, counternormative actions, and nonmoral actions are excluded, and the analysis is conducted solely on the five most established moral foundations, purity is still very distant from the other foundations, which cluster early. See the online supplemental materials for details.

[5] Identical results are obtained if the dissimilarity scores are calculated by subtracting relatedness scores from the maximum observed relatedness (68%) rather than from the maximum possible relatedness (100%). We therefore focus on the conceptually simpler analysis.

[6] The pattern of results is actually somewhat clearer if the dissimilarity scores are treated as interval variables, however, the model stress is unacceptably high under these assumptions (two-dimensional solution: .16; three-dimensional solution: .14). As with the cluster analyses, the results of the MDS analysis remain substantively unchanged if conceptual relatedness is defined as the minimum, maximum, or product of likelihood judgments within a category pair. Also, if the analysis is conducted only on the care, fairness, authority, loyalty, and purity foundations, purity remains very distant from the other foundations. See the online supplemental materials for details.

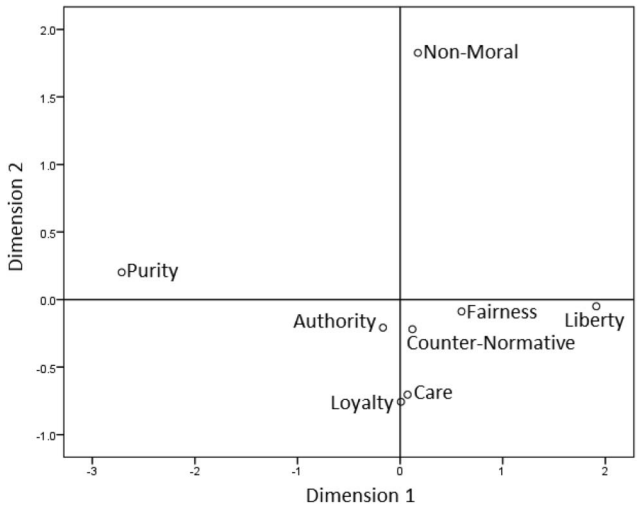

Figure 3. Two-dimensional solution derived from multidimensional scaling of relatedness scores, defined as the mean likelihood judgment within category pairs. Model stress is .08.

and purity violations and nonmoral actions are less closely related. As can be seen in the online supplemental materials, the exact configuration of the eight categories of actions varies slightly depending on how the models are specified (e.g., when conceptual relatedness is defined as the maximum likelihood judgment within a category pair, loyalty violations and counternormative actions cluster first, and liberty violations cluster with nonmoral actions before purity violations do). What is consistent across all of these analyses is that care, fairness, authority, and loyalty violations, and counternormative actions, are always close to one another in space, whereas liberty and purity violations and nonmoral actions are noticeably more distant.

Therefore, we model the virtues of care, fairness, authority, loyalty, and politeness as belonging to a single superordinate category. We cautiously and provisionally label this category *propriety*, which is defined as "conformity to established standards of good or proper behavior or manners" (see http://www.dictionary.com/browse/propriety). The standards of behavior (or, more simply, rules) that these violations offend against are usually explicit (e.g., "respect your elders", "be nice to people"), taught early in life, and relevant to social behavior. We think that this captures what sets these virtues apart from liberty, purity, and nonmoral actions. Liberty is about not *creating* rules that are oppressive or burdensome for others. Purity violations are often highly unusual, so explicit rules forbidding them (e.g., "do not write erotic poetry about your cat") may not be explicitly articulated, or, when rules forbidding less bizarre purity violations are articulated, they tend to be about private or personal conduct, rather than social behavior (e.g., "do not eat pork," "do not have sex outside of marriage"). Last, nonmoral actions (at least, the ones in our stimulus set) have nothing to do with rules at all. Labeling this category *propriety* also accounts for the unexpected finding that counternormative actions clustered with care, fairness, authority, and loyalty violations; the counternormative actions in our stimulus set mostly consist of violations of rules of etiquette that are taught early in life (e.g., "say 'excuse me'"). We should emphasize that this category label is tentative, and that there are not likely to be necessary and sufficient conditions for a violation to be placed in this category. Nevertheless, because we think this provisional label captures the family resemblance among the violations that clustered together, we will use it throughout the remainder of this article. Our derived taxonomy of moral concepts is presented in Figure 4.

This study supports three conclusions. First, our participants possess shared beliefs about how violations of different moral virtues relate. Second, people's representations of moral concepts can be modeled as being organized in a hierarchical, taxonomic structure, and this structure can be uncovered using the methods that we have borrowed from research on category-based induction. Third, this structure is not organized into the social-functional categories of MFT, but it is interpretable, nonetheless, with three distinct categories of violations.

Studies 2a and 2b

We next used two classic methods from the study of category-based induction to test predictions of the taxonomy derived in Study 1 against predictions of an MFT-like taxonomy. In Study 2a, we presented participants with two premises, and asked which one made a conclusion more likely (see López, Atran, Coley, Medin, & Smith, 1997). In Study 2b, we presented participants with one premise, and asked which of two conclusions was more likely, given the premise (see López, 1995; Medin, Lynch, Coley, & Atran, 1997).

Method

Participants. One hundred three participants on Amazon Mechanical Turk began Study 2a, and 100 began Study 2b. After exclusions, we were left with samples of $N = 100$ for both studies (Study 2a: 31 female, $M_{Age} = 34.73$, $SD_{Age} = 9.81$; Study 2b: 41 female, $M_{Age} = 33.62$, $SD_{Age} = 9.02$). The analyses in these studies are one-sample t tests, but we did not have an a priori estimate of the effect sizes we would observe. Based on power analyses conducted with the G*Power software package (Faul, Erdfelder, Lang, & Buchner, 2007), samples of $N = 100$ provide adequate statistical power (.95) to detect a small- to medium-sized effect ($d = .36$). This target sample size was determined before data collection.

Materials and procedure. In these studies, we examined cases where the taxonomy derived in Study 1 and an MFT-like taxonomy make differing predictions about which premises/

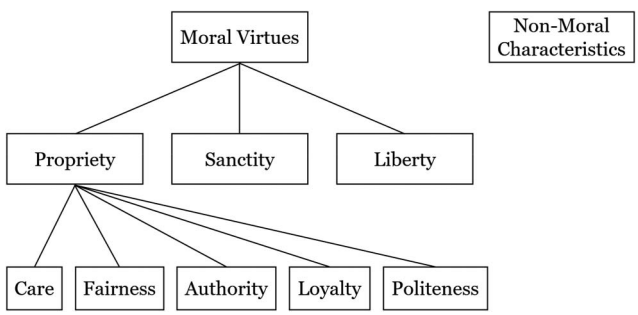

Figure 4. A bottom-up taxonomy of moral concepts derived in Study 1.

conclusions should be more closely related to a given conclusion/premise. For example, in the derived taxonomy, fairness belongs to the same superordinate category as authority, whereas purity does not; in MFT, the reverse is true (see Figures 1 and 4). Therefore, the derived taxonomy predicts that a person who committed a fairness violation should be seen as more likely to commit an authority violation than should someone who committed a purity violation (Study 2a). It similarly predicts that a person who committed an authority violation should be seen as more likely to commit a fairness violation than a purity violation (Study 2b). An MFT-like taxonomy makes the opposite predictions. There are eight such combinations of premises and conclusions, presented in Tables 2 and 3. Politeness and nonmoral characteristics were not included in Studies 2–4, because these studies test the predictions of the derived taxonomy against an MFT-like taxonomy, which would not include them.

In Study 2a, we presented participants with two premises (e.g., "Person A edges out another person in a long line. Person B looks at pornography in which an 18-year-old model has been digitally altered to look like she is 13" [violations of fairness and purity, respectively]), and asked which one more strongly supported a conclusion ("Given this information, which person would be more likely to send out an email to other low-level employees, calling the company president an 'idiot,' if they were a low-level employee?" [authority]). In Study 2b, we presented participants with one premise (e.g., "A low-level company employee sends out an email to other low-level employees, calling the company president an 'idiot.'"), and asked which of two conclusions it more strongly supported ("Given this information, which of the following is this person more likely to do or have done?"). In both studies, participants saw four instances of each of the eight critical premise/conclusion combinations, with all premises and conclusions randomly drawn from the same set of stimuli used in Study 1.[7] The order of presentation of the premises in Study 2a and the conclusions in Study 2b was counterbalanced.

Results

For each participant, we calculated the percentage of responses that agreed with the derived taxonomy (e.g., in Study 2b, the percentage of instances, out of four, when a participant selected a fairness conclusion over a purity conclusion, given an authority premise). The mean percentages for all eight combinations of premises and conclusions were greater than 50% across both studies (significantly so in 15 of 16 cases), and, when averaging across the eight combinations, participants' judgments strongly supported the derived taxonomy (see Tables 2 and 3).[8] Averaging across the eight combinations, 79 out of 100 participants in Study 1a and 89 out of 100 participants in Study 2b made judgments in agreement with the derived taxonomy in more than 50% of cases, significantly greater than 50% of participants, binomial test $ps < .001$. In other words, participants made inductive inferences about immoral behaviors that more closely align with the predictions of the taxonomy derived in Study 1 than those of an MFT-like taxonomy.

Table 2
Descriptive and Inferential Statistics, Study 2a

Conclusion	Premise agreeing with MFT-like taxonomy	Premise agreeing with derived taxonomy	Mean percentage selecting derived	$t(99)$	p	d
Authority	Purity	Care	70.2%	7.75	<.001	.77
Authority	Purity	Fairness	68.0%	4.88	<.001	.49
Loyalty	Purity	Care	60.8%	4.03	<.001	.40
Loyalty	Purity	Fairness	54.3%	1.82	.071	.18
Care	Liberty	Authority	63.7%	5.16	<.001	.52
Care	Liberty	Loyalty	57.0%	2.81	.006	.28
Fairness	Liberty	Authority	58.3%	3.16	.002	.32
Fairness	Liberty	Loyalty	56.3%	2.34	.021	.23
Overall			61.0%	10.34	<.001	1.03

Note. MFT = Moral Foundations Theory.

Table 3
Descriptive and Inferential Statistics, Study 2b

Premise	Conclusion agreeing with MFT-like taxonomy	Conclusion agreeing with derived taxonomy	Mean percentage selecting derived	$t(99)$	p	d
Authority	Purity	Care	86.3%	17.93	<.001	1.79
Authority	Purity	Fairness	85.3%	15.87	<.001	1.59
Loyalty	Purity	Care	78.0%	10.90	<.001	1.09
Loyalty	Purity	Fairness	81.5%	12.85	<.001	1.28
Care	Liberty	Authority	68.0%	6.95	<.001	.70
Care	Liberty	Loyalty	62.3%	4.90	<.001	.49
Fairness	Liberty	Authority	67.5%	6.44	<.001	.64
Fairness	Liberty	Loyalty	59.8%	3.38	.001	.34
Overall			73.6%	16.71	<.001	1.67

Note. MFT = Moral Foundations Theory.

Discussion

These studies used tasks from research on taxonomic mental representations to test the validity of the taxonomy of moral concepts derived in Study 1. Participants consistently made inductive inferences that agreed with the predictions of this model, suggesting that their mental representations of moral concepts resemble this taxonomy more than they do an MFT-like taxonomy.

Study 3

In Study 3, we had participants make likelihood judgments similar to those in Study 1, but used a measure of confirmation as our dependent variable (Tentori, Crupi, Bonini, & Osherson, 2007). Also called "degree of support", confirmation represents how much more or less credible a conclusion becomes, given a premise. This is not the same as the posterior probability of the conclusion, which we examined in Study 1. Rather, it is the contribution made by a premise to the plausibility of a conclusion,

[7] Because of a programming error, not all participants saw four instances of fairness and purity premises and an authority conclusion in Study 2a. For clarity, we present the results as the percentage of responses agreeing with each taxonomy, rather than the raw number.

[8] It might be argued that parametric t tests are inappropriate for this analysis because the variables of interest can only take on five values (0%, 25%, 50%, 75%, 100%), and should not be treated as continuous. The results are the same when Wilcoxon signed-ranks test are used instead.

regardless of the baseline probability of the conclusion being true.[9] Because our stimuli are not equated for baseline frequency (though future research could equate on this dimension, see the online supplemental materials), it was important to test our derived taxonomy using a confirmation measure.

Method

Participants. Four hundred twenty-six participants began the study on Amazon Mechanical Turk. After exclusions, we retained a final sample of 367 (187 female, $M_{Age} = 35.35$, $SD_{Age} = 10.88$). One of the focal tests in this study is a binomial test with a null hypothesis $p_0 = .50$. A sample size of approximately 350 provides good statistical power (.97) to detect a significant difference of .10 (Chow, Shao, & Wang, 2008, pp. 84–85), and we aimed for a sample of approximately this size. This target sample size was determined before data collection.

Materials and procedure. Each participant made 42 total likelihood judgments of the sort examined in Study 1, one for each possible premise/conclusion combination of the six moral foundations, and six baseline judgments with no premise. Premises were randomly selected for each question. Rather than randomly select the conclusion for each question, however, each participant was randomly assigned one of seven conclusions from each foundation, which appeared in all likelihood judgments for that foundation. Each participant saw the same conclusion from every foundation seven times, so that their conditioned judgments were directly comparable to their baseline judgments. Likelihood ratings were made on the same 0% to 100% sliding scale as in Study 1.

Results

We used a simple measure of confirmation, computed by subtracting baseline judgments from conditioned judgments (Eells, 1982; Jeffrey, 1992). For instance, if a participant rated the likelihood of a person driving past an injured man without stopping to help at 30% when they had no other information (i.e., the baseline judgment; care), and the likelihood of this, given that the person had edged out another person in a long line, at 65% (i.e., the conditioned judgment; fairness), the confirmation that this premise brings to this conclusion would be 65% – 30% = 35%.

Both the taxonomy derived in Study 1 and an MFT-like taxonomy classify 18 premise-conclusion pairs as belonging to the same superordinate category (e.g., authority and care are both part of propriety in the derived taxonomy, and loyalty and purity are both binding foundations in MFT), and 18 as belonging to different superordinate categories (e.g., liberty and purity in both taxonomies, see Figures 1 and 4). We calculated the average confirmation gained from premises that belong to the same superordinate category as the conclusion, versus premises that do not, in each taxonomy. We expected that the difference in confirmation between within- and between-category premises would be larger for the derived taxonomy than an MFT-like taxonomy.

Confirmation was substantially higher for within-category premises than between-category premises in both the derived taxonomy ($M_{Within} = 13.31$, $SD_{Within} = 13.65$, $M_{Between} = 3.62$, $SD_{Between} = 11.42$, $t(366) = 16.45$, $p < .001$, $d_{RM} = .86$) and the MFT-like taxonomy ($M_{Within} = 11.60$, $SD_{Within} = 12.06$, $M_{Between} = 5.33$, $SD_{Between} = 11.48$, $t(366) = 17.20$, $p < .001$, $d_{RM} = .90$).[10] This suggests that both taxonomies capture the structure of people's taxonomies of moral concepts to some degree. However, the difference in confirmation between within-category and between-category premises was larger for the derived taxonomy than the MFT-like taxonomy, suggesting that the former is a better model of representations of moral concepts (paired samples t test on the differences between within- and between-category premises: $t(366) = 5.58$, $p < .001$, $d_{RM} = .29$).

We next examined cases where the two taxonomies make differing predictions about which of two premises should be more informative about a given conclusion. These are the same eight combinations of premises and conclusions examined in Study 2a. Paired-samples t tests on the confirmation measure generally agreed with the predictions of the derived taxonomy—in five of eight cases, confirmation was significantly greater given the premise predicted by the derived taxonomy, and in no cases was confirmation significantly greater given the premise predicted by an MFT-like taxonomy (see Table 4). Averaging across the eight premise/conclusion combinations, the results significantly supported the derived taxonomy.

Next, we examined cases where the two taxonomies make differing predictions about which of two conclusions should be more strongly supported by a given premise, analogous to Study 2b. Paired-samples t tests of the confirmation measure were generally directionally consistent with the predictions of the derived taxonomy and not an MFT-like taxonomy, though only significantly so in two cases (see Table 5). Averaging across the eight premise/conclusion combinations, the results significantly supported the derived taxonomy.

Finally, we constructed a vector contrasting the categorizations in the derived taxonomy with those in an MFT-like taxonomy. Premise-conclusion pairs that belong to the same superordinate category in the derived taxonomy but not in an MFT-like taxonomy (e.g., care/authority) were coded as 1, pairs that are in the same superordinate category in an MFT-like taxonomy but not in the derived taxonomy (e.g., care/liberty) were coded as −1, and pairs that both taxonomies categorize in the same way (e.g., care/fairness) were coded as 0. We computed a Pearson correlation between this vector of categorizations and confirmation scores for each participant. A positive correlation indicates that a participant's confirmation scores conform more to the predictions of the derived taxonomy than to those of an MFT-like taxonomy, and a negative correlation indicates the opposite.

Two hundred twenty-six participants out of 366 (62%) expressed a positive correlation (greater than 50%, binomial test $p <$

[9] We thank Daniel Osherson for suggesting that we use a confirmation measure.

[10] d_{RM} denotes the repeated-measures Cohen's d, calculated as the mean within-subjects difference score, divided by the standard deviation of difference scores (see Morris & DeShon, 2002).

.001).[11] Also, the median correlation was significantly larger than 0 (median $r = .044$, $W = 43{,}745$, $p < .001$, mean Fisher-transformed r-to-$z = .048$, one-sample $t(366) = 5.37$, $p < .001$, $d = .28$). Participants' judgments conformed more to the predictions of the derived taxonomy than an MFT-like taxonomy.[12]

Discussion

The confirmation that comes from learning about a prior behavior consistently resembled the predictions of the derived taxonomy more than an MFT-like taxonomy. These results indicate that liberty and purity violations were not rated as less likely in Study 1 and chosen less frequently in Studies 2a and 2b merely because they are seen as statistically rarer. Even accounting for differences in baseline frequency using a confirmation measure, participants' inductive inferences still conformed more to the predictions of the derived taxonomy than an MFT-like taxonomy.

Study 4

In Study 4, we tested cases where the derived taxonomy and an MFT-like taxonomy make differing predictions about which categories are subsumed by a common superordinate category. Participants made judgments about the likelihood of a behavior (conclusion) given two prior behaviors (premises), for premise-premise-conclusion combinations where the two taxonomies differ in their predictions. We predicted that participants' likelihood judgments would conform more to the predictions of the derived taxonomy.

Method

Participants. Four hundred-eighteen participants began the study on Amazon Mechanical Turk. After exclusions, we retained a final sample of 372 (191 female, $M_{Age} = 35.27$, $SD_{Age} = 11.15$). We aimed for a sample of approximately 350, for the same reason as in Study 3. This target sample size was determined before data collection.

Materials and procedure. There are six combinations of two premises and a conclusion for which the two taxonomies make differing predictions about whether the conclusion belongs to the same superordinate category as the premises (see Table 6).

Participants made 24 likelihood judgments of the same form as in Studies 1 and 3, but with two premises instead of one. For example, a participant might learn that a person drove past an injured man on an empty road (care) and edged out another person in a long line (fairness), and then indicate how likely it is that the person would send out an email calling their boss an "idiot" (authority). The premises and conclusions were randomly selected, with the restriction that each participant received four instances of each of the six combinations where the two taxonomies make differing predictions.

Table 4
Within-Subjects t Tests of Confirmation From Different Premises, Study 3

	Within-category premises				
Conclusion	MFT-Like	Derived	$t(366)$	p	d_{RM}
Authority	Purity	Care	5.18	<.001	.27
Authority	Purity	Fairness	.68	.495	.04
Loyalty	Purity	Care	.10	.923	.01
Loyalty	Purity	Fairness	−.61	.540	−.03
Care	Liberty	Authority	4.10	<.001	.21
Care	Liberty	Loyalty	2.15	.032	.11
Fairness	Liberty	Authority	4.11	<.001	.21
Fairness	Liberty	Loyalty	3.56	<.001	.19
Overall			5.09	<.001	.27

Note. MFT = Moral Foundations Theory.

Table 5
Within-Subjects t Tests of Confirmation for Different Conclusions, Study 3

	Within-category Conclusions				
Premise	MFT-Like	Derived	$t(366)$	p	d_{RM}
Authority	Purity	Care	1.70	.091	.09
Authority	Purity	Fairness	1.44	.152	.08
Loyalty	Purity	Care	−.07	.942	−.00
Loyalty	Purity	Fairness	1.26	.208	.07
Care	Liberty	Authority	5.06	<.001	.26
Care	Liberty	Loyalty	1.47	.142	.08
Fairness	Liberty	Authority	2.50	.012	.13
Fairness	Liberty	Loyalty	1.75	.081	.09
Overall			3.68	<.001	.19

Note. MFT = Moral Foundations Theory.

Results

As in Study 3, we created a vector of categorizations derived from the two taxonomies. Premise-premise-conclusion combinations that belong to the same superordinate category in the derived taxonomy, but not an MFT-like taxonomy, were coded as 1, whereas combinations that belong to the same superordinate category in an MFT-like taxonomy, but not the derived taxonomy, were coded as −1. We then computed within-subjects correlations between this vector and participants' likelihood judgments. A positive correlation indicates that a participant's judgments conform more to the predictions of the derived taxonomy than to the predictions of an MFT-like taxonomy, whereas a negative correlation indicates the opposite. Three hundred fifty-seven partici-

[11] One participant responded "50%" to every question, expressing no variance in his judgments. His data were excluded from this analysis.

[12] In Studies 3 and 4, we also computed Anderson's *W,* a statistic that takes each taxonomy's predictions as separate inputs into a mixture model that describes people's judgments. The *W* statistic characterizes the degree to which a participant's responses resemble the predictions of one taxonomy versus the other (see Sanfey & Hastie, 2002). We also separately correlated each participant's responses with the predictions of each taxonomy (same superordinate category = 1, different superordinate category = 0), and examined the percentage of participants with more positive correlations with the derived taxonomy than the MFT-like taxonomy. The results of these analyses were very consistent with the primary analyses (see the online supplemental material, wherein we also report all of these analyses for Study 1).

Table 6
Study 4 Predictions

Premises	Conclusion	Taxonomy predicting same category
Care/Fairness	Liberty	MFT-like
Care/Fairness	Authority	Derived
Care/Fairness	Loyalty	Derived
Authority/Loyalty	Purity	MFT-like
Authority/Loyalty	Care	Derived
Authority/Loyalty	Fairness	Derived

pants (96%) expressed a positive correlation (greater than 50%, binomial test $p < .001$). Also, the median correlation was significantly greater than 0 (median $r = .45$, $W = 69{,}015$, $p < .001$, mean Fisher-transformed r-to-$z = .48$, one-sample $t(371) = 36.15$, $p < .001$, $d = 1.88$).

Discussion

In Study 4, we examined premise-premise-conclusion combinations where an MFT-like taxonomy and the derived taxonomy make different predictions. As in Study 3, participants' judgments conformed more to the predictions of the derived taxonomy than to those of an MFT-like taxonomy.

General Discussion

These studies used the methods of category-based induction research to examine how people represent moral concepts. Five studies suggested that our derived taxonomy reasonably approximates people's representation of the moral concepts under investigation here. Study 1 led us to develop this model based on the convergent results of multiple exploratory analyses. Studies 2 through 4 used confirmatory methods and found that participants' judgments more closely resembled the predictions of the derived taxonomy than the predictions of a taxonomy based on MFT. People seem to have a culturally shared, hierarchical representation of moral concepts that can produce systematic patterns of inductive inference. This taxonomy is not organized into the types of social-functional categories emphasized in MFT, but its structure is interpretable. People seem to view failure to observe rules of proper social conduct (care, fairness, authority, loyalty, and politeness, i.e., propriety), misuse of the body (purity), and oppression (liberty)—at least as they are instantiated in our stimulus set—as distinct categories of violations.

Connection to Past Research

Moral foundations theory. These findings are likely not problematic for MFT, as it is typically articulated. As discussed above, MFT is not usually thought of as a theory of concepts and categories. In MFT, individualizing and binding moral foundations are taken to represent distinct strategies for regulating human behavior, suppressing selfishness, and allowing people to live together cooperatively in groups (Haidt & Kesebir, 2010). Different cultures and subcultures embrace these approaches to different degrees (Haidt, 2012; Haidt & Graham, 2007; Graham et al., 2009)—they elaborate different kinds of "moral systems" (Graham & Haidt, 2010), producing people who have internalized individualizing and binding virtues to greater and lesser extents. Explicit endorsement of the individualizing and binding foundations as morally important virtues supports this assertion; endorsement of care and fairness are highly correlated, as are endorsement of loyalty, authority, and purity, whereas correlations between individualizing and binding foundations (e.g., between care and loyalty) are much lower (Graham et al., 2011; note that liberty had not yet been proposed as a moral foundation when this research was conducted). In short, the individualizing and binding foundations are thought to represent two clusters of virtues that people embrace to varying degrees.

If care, fairness and liberty, and authority, loyalty, and purity form two distinct clusters of virtues that tend to co-occur in people's moral beliefs, then we might expect people to have some knowledge of this. In other words, we would expect people to hold beliefs like "people who care about fairness usually also care about preventing harm" and "people who do not care about loyalty usually also do not care about sexual purity." Consider once again the premise "Joe commits a fairness violation." This premise should indicate to people that Joe does not embrace the individualizing approach to morality, and therefore it should increase the subjective likelihood that Joe would also commit a care violation. It should provide less information, however, about whether Joe embraces the binding approach, and therefore should affect the subjective likelihood that Joe would also commit a loyalty violation to a lesser degree. So, although our results are not problematic for MFT as it is usually articulated, they are inconsistent with a prediction that could reasonably be derived from MFT.

However, what we value in others' behavior and what we expect others to do are different issues. A person might place greater value on rules relating to care and fairness than on rules relating to authority, loyalty, and politeness. But if this person observes another person violate a rule relating to authority, the observer might be unlikely to ignore this potentially valuable information. A key benefit of understanding a person's moral virtues (or lack thereof) is that it aids in predicting their future intentions and behaviors (Cottrell et al., 2007; Landy et al., 2016; Landy & Uhlmann, in press; Pizarro & Tannenbaum, 2011; Wojciszke et al., 1998). The observer might categorize the person who broke a rule related to authority as a "rule-breaker" and use this to predict that this person might be willing to break more valued rules of propriety related to care and fairness. So, although people's explicitly endorsed moral values seem to cluster into individualizing and binding moral foundations, our results suggest that they might predict others' behaviors by assigning people to categories like "rule-breaker," "deviant," or "bully." This could have troubling implications if, for example, an observer sees someone who is unfamiliar with local customs violate an etiquette norm and infer, on this "thin slice" of behavior (Ambady, Bernieri, & Richeson, 2000; Carney, Colvin, & Hall, 2007), that this person might be the type who doesn't care about rules of propriety. This inference could easily affect the way the observer might treat this person, which could, of course, be problematic.

Social domain theory (SDT). Our findings may have implications for SDT (Nucci & Nucci, 1982; Nucci & Turiel, 1978; Smetana, Jambon, & Ball, 2014; Turiel, 1983, 2002, 2014), another prominent theory of moral judgment. SDT states that an action can be wrong in two different senses; moral violations cause

harm or injustice, and are always impermissible, whereas violations of social convention are impermissible only in particular cultures or contexts. The individualizing moral foundations roughly correspond to "moral" wrongs in SDT, with the binding foundations being more similar to "conventional" wrongs. Our results suggest that people might not make the moral/conventional distinction when predicting others' behavior. Some "moral" wrongs (care and fairness violations) are seen, by our participants, as informative about some "conventional" wrongs (authority and loyalty violations) and vice versa. Of course, people may still consider care and fairness violations to be more universally wrong than authority or loyalty violations, as SDT would predict (though see Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993; Landy, 2016; Royzman, Landy, & Goodwin, 2014). Our methods and results cannot speak to this central claim of SDT.

Limitations

Frequency and weirdness. The stimuli used in this study were not equated for their perceived baseline frequency. We collected norming data on this, but it proved impossible to equate for both wrongness and frequency while retaining a sufficiently large number of stimuli to broadly sample each moral foundation. We felt that equating for wrongness was more important, and allowed our stimuli to vary in frequency across the moral foundations. This raises a potential concern that liberty and purity violations may be considered less conceptually related to other types of violations because they are less common, both as premises and as conclusions, in the real world. These types of violations were, in fact, rated as less frequent in our norming studies ($M_{Purity} = 3.97$; $M_{Liberty} = 4.84$; other category Ms > 5.62).

However, differences in frequency cannot fully explain our findings, for at least two reasons. First, we did not find that people's taxonomies of moral concepts divide into "frequent violations" and "infrequent violations." If they had, purity and liberty violations would be close together in people's representations, but, in the MDS analyses in Study 1, they were at least as distant from each other as they were from propriety violations (i.e., care, fairness, authority, loyalty, and etiquette violations). Second, and more importantly, we accounted for differences in baseline frequency using a confirmation measure in Study 3, and found that participants' inductive judgments still more closely resembled the predictions of the derived taxonomy than an MFT-like taxonomy.

Another potential issue is that if purity and liberty violations are perceived as more "weird" than other types of violations, then this difference might partially explain why purity and liberty do not cluster with the others. For two reasons, we do not think that differences in weirdness fully explain the pattern of results we observed. First, we collected norming data on perceived weirdness using a question adapted from Chakroff & Young (2015, Study 2), and ratings of weirdness correlated strongly with ratings of baseline frequency ($r(252) = -.84$ across all 244 stimuli we initially tested, $r(54) = -.86$ across our final stimulus set). Weirdness, as measured here, seems to be largely synonymous with (in)frequency, which we accounted for in Study 3. Second, we ran linear regressions predicting likelihood judgments in Study 1 from frequency/weirdness ratings for each premise and conclusion that a participant saw, and the distance between the premise and the conclusion in the derived taxonomy, with standard errors clustered by participant (to account for repeated judgments within participant). Although the frequency/weirdness of the premise and conclusion significantly affected likelihood judgments, the effects of distance were also significant, indicating that the structure of the derived taxonomy accounts for variance in judgments over and above baseline frequency or weirdness (see the online supplemental materials for details).

That said, baseline frequency and weirdness surely do contribute to likelihood judgments of moral violations (see, e.g., Gray & Keeney, 2015), and a full model of how these judgments are made would incorporate information about the prior probabilities of different acts. Such a model could include terms describing both the relatedness and/or causal connections between the premises and conclusions as well as their prior probabilities (see, e.g., Blok, Osherson, & Medin, 2007; Rehder, 2017). Developing and testing a formal model of moral induction is a fruitful direction for future research.

Stimulus selection. We used stimuli that were pretested to be uniquely good exemplars of the six moral foundations, and we tried to broadly cover each moral foundation. However, we cannot be sure that our stimuli constitute a representative sample of violations of each foundation (if such a thing is even possible). Radically different stimuli could potentially produce different results. Moreover, we may have omitted other virtues that feature prominently in people's representations of moral concepts. We think that an especially strong candidate for such an overlooked virtue is honesty, which has emerged as a central virtue in several distinct projects aimed at understanding lay conceptions of moral character (Aquino & Reed, 2002; Lapsley & Lasky, 2001; Walker & Hennig, 2004; Walker & Pitts, 1998). Although honesty is not emphasized in MFT (though see Iyer, 2010), there is evidence that it holds an important place in people's understanding of morality. For instance, knowledge of a person's honesty is rated as highly useful for understanding their moral character (Goodwin, Piazza, & Rozin, 2014), honesty loads with other moral traits in exploratory factor analyses of ratings of real people's personalities (Landy et al., 2016) and honesty is considered to be among the most fundamental elements in one's identity (Strohminger & Nichols, 2014). Therefore, our derived taxonomy may not fully characterize people's theories of morality. Future research that does not take any particular theory of morality as a starting point could inform this issue. Developing a fully bottom-up model of moral concepts that does not inherit assumptions from any prior theory is a difficult, but important, task for future research.

Future Directions

Other methods. The current studies have used people's judgments about the likelihood of behaviors to model their taxonomies of moral concepts, and this is just one of many paradigms that could be used to examine the relatedness of concepts (see Balota & Coane, 2008). Some other paradigms that have proven useful for this purpose, historically, have analyzed reaction times (RTs) (dating back at least to Collins & Quillian, 1969). Although our stimuli were normed on several characteristics (and equated on mean levels of moral wrongness; see the five norming studies in the online supplemental

material), they differ on other characteristics (e.g., length, word frequency, co-occurrence of words in common contexts, and reading level) that would complicate analyses of RTs (see Clifford et al., 2015, for stimuli that are more closely equated on these sorts of low-level features). That said, we think that some predictions about RTs could be motivated by our framework. For example, if a participant has just judged an authority violation, she might be faster to judge a subsequent fairness violation than a subsequent purity violation, all else being equal, because if authority and fairness violations belong to the same superordinate category, accessing one might increase the accessibility of the other. Although we are aware of several complications in interpreting RTs in paradigms like these (see, e.g., Hutchison, 2003; Lucas, 2000), we are cautiously optimistic that testing predictions about RTs derived from a framework like ours is a potentially informative direction for future research.

Other people. Exploring how other cultures represent moral concepts is an important avenue for future work (Henrich, Heine, & Norenzayan, 2010). Although the cultural consensus analysis in Study 1 suggests that our sample drew on shared conceptual knowledge of how different kinds of violations relate, it is possible that these methods could yield different taxonomies for other cultures. For example, in cultures where purity rules are more strongly emphasized and more integrated into everyday behavior, purity might be more related to propriety than it was in our samples. Moreover, exploring how different subcultures represent moral concepts is also a potentially fruitful task. Even though the Cultural Consensus Analysis indicates that our American participants agree on how different kinds of violations relate, to a first approximation, there could still be meaningful subgroup differences (see, e.g., Medin et al., 1997). We have no reason to assume any particular differences a priori, but it is possible that there might be differences in how different groups represent these concepts.

As an initial test of this possibility, we regressed the within-subjects correlations from Studies 1, 3, and 4, and the average percentage of choices agreeing with the derived taxonomy in Studies 2a and 2b on the demographic information we collected (age, sex, race/ethnicity, political views, and education). These demographic characteristics generally proved to be relatively weak predictors of whether people represented moral concepts in a way that is more similar to the derived or MFT-like taxonomy (i.e., the small number of significant relationships that were detected [three out of 40 tests] does not differ from what would be expected by chance alone, assuming a Type I error rate of .05, binomial test $p = .138$, and were inconsistent across studies). That said, further exploring such between-groups differences, when they occur, could further illuminate how people represent moral actions and violations.

Conclusion

In conclusion, we derived a taxonomic model of how people mentally represent moral concepts that is organized by three superordinate categories of virtues: propriety, purity, and liberty. Although this taxonomic structure does not resemble the structure of Moral Foundations Theory, it is interpretable. This research provides new insight into how people parse their moral worlds, and demonstrates how methods from category-based induction research can be used to test the psychological plausibility of theories of moral judgment. More broadly, we think this research illustrates how classic methods from the psychology of concepts and categories can be used to answer questions in other areas of research, such as moral psychology and social psychology.

Context of the Research

The motivation for this research is based on Haidt and Joseph's (2004) initial formulation of Moral Foundations Theory, which presents it as a conceptual scheme for understanding the pan-culturally recognized virtues that form the basis of most moral codes. We realized that relatively little was known about how well this conceptual scheme matches the one that people carry around in their heads, and that this could be tested using classic methods from cognitive psychology. The first author's research is largely about judgments of moral violations and virtues, and the second author has long-standing research interests in both moral judgment and decision making and in concepts and categories, making this project a natural extension of both of our wider programs of research.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32,* 201–271. http://dx.doi.org/10.1016/S0065-2601(00)80006-4

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

Balota, D. A., & Coane, J. H. (2008). Semantic memory. In J. H. Byrne, H. Eichenbaum, R. Mwenzel, H. L. Roediger, III, & D. Sweatt (Eds.), *Learning and memory: A comprehensive reference* (pp. 511–534). Amsterdam, the Netherlands: Elsevier. http://dx.doi.org/10.1016/B978-012370509-9.00170-4

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2016). Moral judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 478–515). Chichester, UK: Wiley.

Blok, S., Osherson, D., & Medin, D. L. (2007). From similarity to chance. In A. Feeney & E. Heit (Eds.), *Inductive reasoning*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511619304.007

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Bradford Books.

Carney, D. R., Colvin, R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41,* 1054–1072. http://dx.doi.org/10.1016/j.jrp.2007.01.004

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PLoS ONE, 8,* e74434. http://dx.doi.org/10.1371/journal.pone.0074434

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition, 136,* 30–37. http://dx.doi.org/10.1016/j.cognition.2014.11.034

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition, 65,* 15–32. http://dx.doi.org/10.1016/S0010-0277(97)00034-6

Chow, S.-C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research*. Boca Raton, LA: Chapman & Hall.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods, 47,* 1178–1198. http://dx.doi.org/10.3758/s13428-014-0551-2

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–247. http://dx.doi.org/10.1016/S0022-5371(69)80069-1

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology, 92,* 208–231. http://dx.doi.org/10.1037/0022-3514.92.2.208

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin, 40,* 1559–1573. http://dx.doi.org/10.1177/0146167214551152

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS ONE, 12,* e0173405. http://dx.doi.org/10.1371/journal.pone.0173405

Eells, E. (1982). *Rational decision and causality*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781316534823

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Feeney, A., & Heit, E. (2007). *Inductive reasoning*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511619304

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26,* 796–804. http://dx.doi.org/10.1037/0012-1649.26.5.796

Glenn, A. L., Iyer, R., Graham, J., Koleva, S., & Haidt, J. (2009). Are all types of morality compromised in psychopathy? *Journal of Personality Disorders, 23,* 384–398. http://dx.doi.org/10.1521/pedi.2009.23.4.384

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. http://dx.doi.org/10.1037/a0034726

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101,* 366–385.

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review, 14,* 140–150. http://dx.doi.org/10.1177/1088868309353415

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil*. New York, NY: APA Books. http://dx.doi.org/10.1037/13091-001

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029–1046. http://dx.doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLoS ONE, 7,* e50092. http://dx.doi.org/10.1371/journal.pone.0050092

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological & Personality Science, 6,* 859–868. http://dx.doi.org/10.1177/1948550615592241

Haidt, J. (2008). Morality. *Perspectives on Psychological Science, 3,* 65–72. http://dx.doi.org/10.1111/j.1745-6916.2008.00063.x

Haidt, J. (2012). *The righteous mind*. New York, NY: Vintage Books.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). *Moral dumbfounding: When intuition finds no reason*. Unpublished manuscript, Department of Psychology, University of Virginia, Charlottesville, VA.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20,* 98–116. http://dx.doi.org/10.1007/s11211-007-0034-z

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31,* 191–221. http://dx.doi.org/10.1111/j.1559-1816.2001.tb02489.x

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133,* 55–66. http://dx.doi.org/10.1162/0011526042365555

Haidt, J., & Kesebir, S. (2010). Morality. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 797–832). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9780470561119.socpsy002022

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613–628. http://dx.doi.org/10.1037/0022-3514.65.4.613

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569–592. http://dx.doi.org/10.3758/BF03212996

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33,* 111–135. http://dx.doi.org/10.1017/S0140525X10000725

Huebner, B., Lee, J. J., & Hauser, M. D. (2010). The moral-conventional distinction in mature moral competence. *Journal of Cognition and Culture, 10,* 1–26. http://dx.doi.org/10.1163/156853710X497149

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10,* 785–813. http://dx.doi.org/10.3758/BF03196544

Iyer, R. (2010, December 7). *The case for honesty as a moral foundation*. Retrieved from http://www.polipsych.com/2010/12/07/the-case-for-honesty-as-a-moral-foundation/

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE, 7,* e42366. http://dx.doi.org/10.1371/journal.pone.0042366

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review, 17,* 219–236. http://dx.doi.org/10.1177/1088868313480274

Jeffrey, R. (1992). *Probability and the art of judgment*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139172394

Kruskal, J. B. (1964a). Nonmetric multidimensional scaling: I. *Psychometrika, 29,* 1–27. http://dx.doi.org/10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: II. *Psychometrika, 29,* 115–129. http://dx.doi.org/10.1007/BF02289694

Landy, J. F. (2016). Representations of moral violations: Category members and associated features. *Judgment and Decision Making, 11,* 496–508.

Landy, J. F., & Piazza, J. (in press). Re-evaluating moral disgust: Sensitivity to many affective states predicts extremity in many evaluative judgments. *Social Psychological & Personality Science*.

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. http://dx.doi.org/10.1177/0146167216655984

Landy, J. F., & Uhlmann, E. L. (in press). Morality is personal. In K. Gray & J. Graham (Eds.), *The atlas of moral psychology: Mapping good and evil in the mind*. New York, NY: Guilford Press.

Lapsley, D. K., & Lasky, B. (2001). Prototypic moral character. *Identity: An International Journal of Theory and Research, 1,* 345–363. http://dx.doi.org/10.1207/S1532706XID0104_03

López, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition, 23,* 374–382. http://dx.doi.org/10.3758/BF03197238

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32,* 251–295. http://dx.doi.org/10.1006/cogp.1997.0651

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7,* 618–630. http://dx.doi.org/10.3758/BF03212999

McAdams, D. P., Albaugh, M., Farber, E., Daniels, J., Logan, R. L., & Olson, B. (2008). Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of Personality and Social Psychology, 95,* 978–990. http://dx.doi.org/10.1037/a0012650

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111,* 960–983. http://dx.doi.org/10.1037/0033-295X.111.4.960

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49–96. http://dx.doi.org/10.1006/cogp.1997.0645

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition, 99,* 237–273. http://dx.doi.org/10.1016/j.cognition.2003.12.005

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105–125. http://dx.doi.org/10.1037/1082-989X.7.1.105

Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin, 42,* 1227–1242. http://dx.doi.org/10.1177/0146167216653933

Nucci, L. P., & Nucci, M. S. (1982). Children's responses to moral and social conventional transgressions in free-play settings. *Child Development, 53,* 1337–1342. http://dx.doi.org/10.2307/1129024

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49,* 400–407. http://dx.doi.org/10.2307/1128704

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185–200. http://dx.doi.org/10.1037/0033-295X.97.2.185

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience, 23,* 3162–3180. http://dx.doi.org/10.1162/jocn_a_00017

Piazza, J., & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making, 8,* 639–661.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Rehder, B. (2017). Concepts as causal models: Induction. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 377–413). New York, NY: Oxford University Press.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin, 127,* 827–852. http://dx.doi.org/10.1037/0033-2909.127.6.827

Rips, L. J., Smith, E. E., & Medin, D. L. (2012). Concepts and categories: Memory, meaning, and metaphysics. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (177–209). New York, NY: New York: Oxford University Press.

Robitzsch, A., Grund, S., & Henke, T. (2017, August 24). *Package 'miceadds'*. Retrieved from https://cran.r-project.org/web/packages/miceadds/miceadds.pdf

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus. *American Behavioral Scientist, 31,* 163–177. http://dx.doi.org/10.1177/000276487031002003

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9,* 283–294. http://dx.doi.org/10.1037/h0026086

Royzman, E., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion, 14,* 892–907. http://dx.doi.org/10.1037/a0036829

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making, 9,* 176–190.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition, 112,* 159–174. http://dx.doi.org/10.1016/j.cognition.2009.04.004

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76,* 574–586. http://dx.doi.org/10.1037/0022-3514.76.4.574

Sanfey, A. G., & Hastie, R. (2002). Interevent relationships and judgment under uncertainty: Structure determines strategy. *Memory & Cognition, 30,* 921–933. http://dx.doi.org/10.3758/BF03195777

Schier, U. K., Ockenfels, A., & Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology, 56,* 107–115. http://dx.doi.org/10.1016/j.joep.2016.06.004

Sinnott-Armstrong, W. (Ed.). (2008). *The cognitive science of morality: Intuition and diversity*. Cambridge, MA: MIT Press.

Smetana, J. G., Jambon, M., & Ball, C. (2014). The social domain approach to children's moral and social judgments. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 23–45). New York, NY: Psychology Press. http://dx.doi.org/10.4324/9780203581957.ch2

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Science Bulletin, 38,* 1409–1438.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131,* 159–171. http://dx.doi.org/10.1016/j.cognition.2013.12.005

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition, 103,* 107–119. http://dx.doi.org/10.1016/j.cognition.2005.09.006

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. New York, NY: Cambridge University Press.

Turiel, E. (2002). *The culture of morality*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511613500

Turiel, E. (2014). Morality: Epistemology, development, and social opposition. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 3–22). New York, NY: Psychology Press. http://dx.doi.org/10.4324/9780203581957.ch1

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology, 86,* 629–647.

Walker, L. J., & Pitts, R. C. (1998). Naturalistic conceptions of moral maturity. *Developmental Psychology, 34,* 403–419. http://dx.doi.org/10.1037/0012-1649.34.3.403

Walker, L. J., & Pitts, R. C. (1998). Naturalistic conceptions of moral maturity. *Developmental Psychology, 34,* 403–419. http://dx.doi.org/10.1037/0012-1649.34.3.403

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field Methods, 19,* 339–368. http://dx.doi.org/10.1177/1525822X07303502

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

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology, 11,* 8–21. http://dx.doi.org/10.20982/tqmp.11.1.p008

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120,* 202–214. http://dx.doi.org/10.1016/j.cognition.2011.04.005

Received July 19, 2017
Revision received December 6, 2017
Accepted December 7, 2017 ■

Supplemental Materials:

An Empirically-Derived Taxonomy of Moral Concepts

Justin F. Landy & Daniel M. Bartels

University of Chicago

Norming Study 1

Method

Participants. Three hundred-five adults located in the United States were recruited through Amazon Mechanical Turk. After exclusions, we retained a final sample of $N = 288$ (137 female).

Stimuli. Two hundred forty-four descriptions of behaviors were constructed, with the intention of covering (i) a wide range of violations of all six moral foundations, as well as (ii) violations of social norms or conventions that do not clearly violate any of the six foundations, and (iii) a variety of morally irrelevant actions. Many of the described behaviors were adapted from previous research on MFT and moral judgment more generally (Clifford et al., 2015; Graham & Haidt, 2012; Haidt, Bjorklund, & Murphy, 2000; Haidt, Koller, & Dias, 1993; Huebner, Lee & Hauser, 2010; Landy & Piazza, in press; Parkinson et al., 2011; Royzman, Atanasov, Landy, Parks, & Gepty, 2014; Royzman, Leeman, & Baron, 2009; Rozin, Lowery, Imada, & Haidt, 1999; Turiel, 1983), while others were original. All adapted stimuli were rewritten so that they shared a common form – 1-2 sentence, present-tense descriptions of a single, unnamed target person engaging in a specified behavior. This often necessitated shortening existing stimuli, sometimes by a substantial amount. We strove to maintain the meaning of all adapted stimuli while making them brief and easy to read. We also modified stimuli to avoid any reference to specific nations, religious groups, or other social groups, reasoning that we might use these stimuli in the future for cross-cultural research. We predicted that the stimuli would fall naturally into eight conceptual categories: violations of the foundations of care, fairness, liberty, loyalty, authority, and purity, miscellaneous counter-normative actions that do not represent violations of any foundation, and non-moral acts. The initial stimulus set included at least 25 behaviors that we expected to fall into each of the eight categories.

Procedure. After consenting to participate, participants were randomly assigned to answer one of four questions – how wrong the described action is, how often such actions occur in the world, how "weird (i.e., unusual, bizarre, odd)" the action is (adapted from Chakroff & Young, 2015, Study 2), and how much "someone was negatively affected by this action (physically, psychologically, emotionally, materially, or otherwise), other than the person that did it" (adapted from Royzman et al., 2009). This latter question was intended to tap perceptions of how much harm an action caused, in a broad sense; perceptions of the narrower sense of the term that defines the care/harm foundation were assessed in the next study. We focus here on responses to the first (wrongness) question. Full norming data are available by request.

Participants each viewed 120 randomly-selected behavioral descriptions and answered their randomly-assigned question for each one on 1-9 scales. After responding to 120 stimuli, participants filled out a brief demographics questionnaire, and were debriefed, thanked, and paid.

Results

An average of 34 participants rated each of the 244 stimuli on each of the four questions (range: 24-50). Mean wrongness ratings are presented in Table S2.

Norming Study 2

Method

Participants. Two hundred seventy-six adults located in the United States who did not participate in the prior study were recruited through Amazon Mechanical Turk. After exclusons, we retained a final sample of $N = 257$ (124 female).

Stimuli. The same 244 stimuli were used in this study.

Procedure. After consenting to participate, participants were presented with 30 randomly-selected behavioral descriptions. For each description, participants indicated their agreement with seven statements on nine-point scales (1 = "Strongly disagree"; 9 = "Strongly agree"). Six of these statements indicated that the described action exemplified one of the six moral foundations. An analogous seventh question was also included, tapping how much the action violated principles of honesty. This question was included for exploratory purposes and is not analyzed further here. The seven dependent variables are presented in Table S1. Notably, our question format allowed

participants to rate a given behavior as exemplifying one moral foundation, many, or none, providing us with information about which behaviors are more or less "pure" instantiations of each foundation.

Table 1. Statements used in Norming Studies 2, 3a, and 3b to define the six moral foundations and honesty.

Foundation	Study 2 Statement	Study 3 Statement
Care/harm	This action was harmful, damaging, or caused suffering to someone other than the person that did it.	This action is harmful, cruel, and/or caused suffering.
Fairness/reciprocity	This action was unfair, unjust, or violated someone's rights.	This action is unfair, unjust, and/or an example of cheating at something.
Liberty/oppression	This action was an attempt to dominate someone, restrict their liberty, or deny them their freedom.	This action is an attempt to dominate someone, restrict their liberty, and/or oppress them.
Loyalty/betrayal	This action was disloyal, treacherous, or betrayed an important group.	This action is disloyal to an important group, treacherous, and/or an act of betrayal.
Authority/subversion	This action was disrespectful of authority, disobedient, or rebellious.	This action is disrespectful of authority, disobedient, and/or subversive.
Purity/degradation	This action was disgusting, impure, or obscene.	This action is disgusting, impure, and/or degrading.
Honesty	This action was dishonest, deceitful, or deceptive.	N/A
Non-Moral	N/A	None of the above – there is nothing wrong with this action.
Miscellaneous Counter-Normative	N/A	None of the above, but this action is still something that people should not do. (Study 3b only)

After responding to 30 behaviors, participants completed a brief demographics questionnaire, and were debriefed, thanked, and paid.

Results

An average of 31 participants rated each of the 244 stimuli (range: 19-43). We present the foundation that each scenario was rated as most exemplifying, as well as the mean rating for that foundation (the "foundation score") in Table S2. Data from this study were used to create a list of candidates for our final stimulus set. These candidates were then evaluated using a new procedure in Norming Studies 3a and 3b.

Moral Violations. To be considered a candidate exemplar of a given moral foundation, a stimulus had to (i) be rated as better exemplifying that foundation than any other, and (ii) have a mean rating of at least the scale midpoint on that foundation—the stimulus had to be rated as at least a moderately good exemplar of at least one moral foundation. We intentionally chose this liberal criterion for two reasons. First, Norming Study 3a will serve as a confirmatory analysis, excluding any candidates that do not belong in a given category. Second, we wanted our set of candidate stimuli to be quite large and diverse, to provide broad conceptual coverage of the moral foundations (we did not want all of our candidate purity offenses to be sexual in nature, for example). One hundred ninety-three total candidate moral violations were included in Norming Study 3a based on this criterion, some of which were originally intended as miscellaneous counter-normative actions.

Counter-normative Actions. Over half of our hypothesized counter-normative behaviors (13 out of 25) scored above the midpoint on at least one foundation rating, making them candidate exemplars of moral violations, rather than miscellaneous counter-normative acts. The remaining 11 were mostly mild violations of etiquette (e.g., failing to send a thank-you letter after a job interview, attending a pot luck dinner without bringing any food). This is a much narrower set of behaviors than what Turiel (1983) meant by "conventional violations," and are probably better conceptualized as faux pas. Once loyalty, authority, and purity violations are removed, relatively little remains in our set of social conventional violations. Nonetheless, we ran a confirmatory study to validate as many of these actions as we could (Norming Study 3b).

Non-Moral Actions. Twenty-four out of 25 actions that were hypothesized to be devoid of moral content were rated below the midpoint on all six foundation scales. The sole exception was a scenario about a man using a dead squid as a puppet, which was seen as resembling a purity violation (M = 5.71). We retained the 24 squidless stimuli for Norming Studies 3a and 3b. As would be expected, these stimuli were rated very low on wrongness in Norming Study 1 (M = 1.77, on a 1-9 scale), indicating that they are indeed morally inert.

Norming Studies 3a and 3b

One hundred ninety-three candidate exemplars of the moral foundations, 24 candidate non-moral actions, and 11 candidate miscellaneous counter-normative actions were retained from the previous study. Not all candidates were rated most highly on the foundation that we originally hypothesized that they would exemplify. For instance, "A manager at a restaurant requires all of her

servers to split their tips evenly, even though some earn larger tips than others" was originally intended as a less politically charged analog for taxation and wealth redistribution, and, therefore, as a liberty violation (see Haidt, 2012), but was rated as most violating principles of fairness. We therefore wanted to test whether there was widespread agreement about which foundation each candidate stimulus best exemplifies. Thus, in Norming Studies 3a and 3b we had participants engage in a choice task, rather than a rating task, and assign each candidate behavior to the foundation that it best exemplified.

Method

Participants. Norming Study 3a was run in three waves, with each candidate moral violation and non-moral action included in one of the three.[13] Four hundred eighty-one total participants were recruited through Amazon Mechanical Turk. After exclusions, we retained a final sample of N = 464. Norming Study 3b was run in a single wave, and included all hypothesized non-moral and counter-normative actions. One hundred-eleven participants were recruited through Amazon Mechanical Turk. After exclusions, we retained a final sample of N = 108. No participant took part in more than one wave of the studies, or had participated in Norming Study 1 or 2.

Norming Study 3a Procedure. In waves 1 and 3, participants were presented with a random subset of 20 stimuli. In wave 2, all participants were presented with the same 11 stimuli. In all three waves, participants were asked to indicate which of seven statements best described the behavior (see Table S1). Six of these statements defined the moral foundations, and the seventh indicated that there was "nothing wrong with" the action. The definitions of the moral foundations were modified somewhat from Norming Study 2, primarily to ensure that congruence in responses across the studies is not due to some idiosyncratic property of the way we defined the moral foundations. Notably, every definition in this study included a form of the negative pole of the foundation name (e.g., "cheating", "subversive"). Most also included some negation of the positive pole of the foundation name as well (e.g., "unfair", "disrespectful of authority"), though *care* and *purity* did not, the former because it was linguistically awkward, and the latter because we thought that the term would be too esoteric for participants. After responding all of their assigned stimuli, participants completed a brief demographics questionnaire, and were debriefed, thanked, and paid.

[13] Two stimuli were repeated between waves 1 and 2. We report only the wave 1 ratings for simplicity; the wave 2 ratings were similar.

Norming Study 3b Procedure. Participants were first presented with one exemplar of each moral foundation that was validated in Norming Study 3a to calibrate their use of the choice options. They were then presented with a random subset of 20 actions originally hypothesized to be non-moral or counter-normative, exemplifying no particular moral foundation. As in Norming Study 3a, participants selected the statement that best described each presented action, but an eighth choice was included that indicated that the action did not fit any of the above definitions, but still should not be done (i.e., that it is counter-normative, but does not exemplify any moral foundation, see Table S1). After responding to all of their assigned stimuli, participants completed a brief demographics questionnaire, and were debriefed, thanked, and paid.

Results

Moral Violations. A candidate stimulus was considered validated if at least 50% of participants in Norming Study 3a confirmed the tentative classification based on the ratings in Norming Study 2. One hundred thirty-three moral violations were retained based on this criterion.

Non-Moral Actions. At least 50% of participants in Norming Study 3b indicated that there was "nothing wrong with" 19 out of 24 non-moral actions retained from Norming Study 2. This proved to be a slightly more conservative test of these stimuli than was Norming Study 3a, where participants indicated that there was "nothing wrong with" 23 of the actions. We therefore retained the 19 actions from Norming Study 3b.

Counter-Normative Actions. Ten actions in Norming Study 3b were classified as not fitting the definition of any moral foundation, but still being things that people should not do. One of these actions had an authority foundation score greater than five in Norming Study 1, and was therefore considered a possible authority violation, rather than a miscellaneous counter-normative act—in fact, it had already been retained in our final stimulus set on the basis of Norming Study 3a. We retained the other nine actions.

Norming Study 4

For this final norming study, we pared down the 161 stimuli retained from the prior studies to produce an early version of the final stimulus set used in the studies reported in the main text. We then modified them to remove any gender information, and confirmed that this did not substantively

impact the psychometric properties of the stimuli. The final stimulus set used in the main studies is this "genderless" set, and is presented in Table S2.

Method

Stimulus selection. As noted in the main text, we felt that it was important to equate the exemplars of each moral foundation in their moral wrongness. Wrongness ratings from Norming Study 1 were used to select a subset of the actions retained from the prior studies. We aimed to select stimuli that were good exemplars of each moral foundation (i.e., that were rated as representing a foundation well in Norming Study 2, and frequently categorized as representing that foundation in Norming Study 3a) that also gave us broad conceptual coverage of the foundation (e.g., our Purity stimuli cover food taboos, deviant sex, promiscuity, and animalistic behavior), all while maintaining similar mean wrongness scores across the six foundations. We did not strictly prioritize any one of these three (often competing) considerations, but rather tried to balance them. The final stimulus set consists of seven exemplars of each moral foundation, seven miscellaneous counter-normative actions, and seven non-moral actions.

The six foundations are closely equated in their moral wrongness (range: 5.20-5.26). In all but one case, the moral violations conform to our a priori classification (the one exception is the tip-splitting scenario mentioned above, which we considered a liberty violation a priori, but was consistently rated as violating principles of fairness), indicating that both experts and lay people view them as good exemplars of their moral foundations. Also, these moral violations are *distinctly* good exemplars of their respective foundations. The scenarios' highest foundation scores in Norming Study 2 were, on average, 2.49 scale points (out of nine) higher than the average of their other five foundation scores, and 1.28 scale points higher than their second highest foundation scores (mean repeated-measures *d*s 1.26 and .30). These scenarios not only *best* exemplify one foundation, they exemplify one foundation *substantially better* than any other. Qualitatively, the stimuli provide broad conceptual coverage of each foundation (e.g., the care violations include both physical and emotional harms, the liberty violations include both overbearing parents and overreaching politicians, etc.). Lastly, the non-moral acts were rated as extremely low on moral wrongness (M = 1.15).

Participants. One hundred fifty-six adults located in the United States were recruited through Amazon Mechanical Turk. MTurk workers who had participated in any of the previous

studies could not participate in Norming Study 4. After exclusions, we retained a final sample of N = 150 (67 female).

Procedure. After giving informed consent, participants were randomly assigned to one of three conditions: male, female, or genderless. They were then presented with 48 behavioral descriptions, in a randomized order (eight scenarios where the target's gender is never mentioned were excluded, e.g., "A staff member talks loudly and interrupts the mayor's speech to the public"). The scenarios were modified such that, depending on condition, the people described in the scenarios were either male, female, or were described as "a person", without reference to gender. In some cases, the content of the scenario had to be slightly modified to retain the intended meaning – for instance, a scenario about wearing a favorite dress was changed to a favorite suit in the male condition, and a favorite outfit in the genderless condition. After each description, participants indicated how wrong the action in question was. After responding to all 48 scenarios, participants completed a brief demographics questionnaire and were debriefed, thanked, and paid.

Results

We compared the wrongness ratings of each of the 48 presented scenarios across the three conditions. Out of 144 pairwise comparisons between conditions, only four were significantly different by between-subjects t-tests before correcting for multiple comparisons. After applying a Bonferroni correction, none of these comparisons were significant. Modifying the gender of the characters in our stimuli did not meaningfully affect wrongness ratings. We therefore used the "genderless" stimuli in our main studies. The final stimulus set is presented in Table S2.

Table S2. Final, "genderless" stimulus set, with selected norming data.

Scenario	Source	Hypothesized Foundation	Empirical Foundation	Foundation Score	%	Wrong
Care/Harm Scenarios (Mean Wrongness: 5.24)						
A young person shoots geese repeatedly with a pellet gun out in the woods.	Clifford et al. (2015)	Care	Care	7.30	84%	6.00
A person spanks their child with a spatula for getting bad grades in school.	Clifford et al. (2015)	Care	Care	6.09	55%	5.43
A person makes fun of their brother for getting dumped by his girlfriend.	Clifford et al. (2015)	Care	Care	7.03	63%	5.03
A person tells a woman that her painting looks like it was done by children.	Clifford et al. (2015)	Care	Care	6.40	53%	4.55
A young person tells a boy that his older brother is much more attractive than him.	Clifford et al. (2015)	Care	Care	6.48	53%	3.97
A person is driving along an empty road	Original	Care	Care	6.87	78%	6.00

Scenario	Source	Hypothesized Foundation	Empirical Foundation	Foundation Score	%	Wrong
and sees a man who is clearly injured. They drive past the man and do not stop to help him.						
A young person laughs when they realize their friend's dad is a janitor.	Clifford et al. (2015)	Care	Care	6.88	58%	5.73
Fairness/Cheating Scenarios (Mean Wrongness: 5.22)						
A person skips to the front of the line because their friend is an employee.	Clifford et al. (2015)	Fairness	Fairness	7.19	71%	5.49
A soccer player pretends to be seriously fouled by an opposing player during a local game.	Clifford et al. (2015)	Fairness	Fairness	6.30	88%	5.17
A person hires their nephew for a job, instead of a more qualified applicant.	Original	Fairness	Fairness	6.25	68%	4.94
A tenant bribes a landlord to be the first to get their apartment repainted.	Clifford et al. (2015)	Fairness	Fairness	5.47	85%	4.89
A person edges out another person in a long line.	Rozin et al. (1999)	Fairness	Fairness	6.37	78%	4.70
A runner takes a shortcut on the course during the marathon in order to win.	Clifford et al. (2015)	Fairness	Fairness	7.46	87%	6.45
A manager at a restaurant requires all of their servers to split their tips evenly, even though some earn larger tips than others.	Original	Liberty	Fairness	5.83	51%	4.93
Liberty/Oppression Scenarios (Mean Wrongness: 5.23)						
A mayor imposes a curfew on all residents of their town.	Original	Liberty	Liberty	7.72	74%	5.74
A public leader tries to ban the wearing of hooded sweatshirts.	Clifford et al. (2015)	Liberty	Liberty	7.36	65%	5.65
A person tells their fiance that they have to switch to the person's political party.	Clifford et al. (2015)	Liberty	Liberty	7.57	70%	5.32
A person pressures their daughter to become a famous evening news anchor.	Clifford et al. (2015)	Liberty	Liberty	6.38	73%	4.12
A religious leader bans their congregants from wearing bright colors during worship.	Clifford et al. (2015)	Liberty	Liberty	7.10	84%	4.11
A parent forces their daughter to enroll as a pre-med student in college.	Clifford et al. (2015)	Liberty	Liberty	7.59	86%	5.51
A person tells their romantic partner that they must convert to the person's religion.	Clifford et al. (2015)	Liberty	Liberty	7.15	68%	6.17
Loyalty/Betrayal Scenarios (Mean Wrongness: 5.26)						
A person sells government secrets from their nation to another country. The information is classified, though the sale will not have any negative consequences for people in the person's home nation.	Original	Loyalty	Loyalty	7.94	92%	6.33
An ambassador jokes about the stupidity of people from their home country.	Clifford et al. (2015)	Loyalty	Loyalty	6.30	56%	5.06
A person shares a secret that their best friend entrusted to them. The secret is not damaging or embarrassing in any way and the friend never finds out that their secret has been shared.	Original	Loyalty	Loyalty	6.88	69%	5.00
A coach celebrates with the opposing team's players who just won the game.	Clifford et al. (2015)	Loyalty	Loyalty	6.78	61%	4.54

Scenario	Source	Hypothesized Foundation	Empirical Foundation	Foundation Score	%	Wrong
A person leaves their family business to go work for their main competitor.	Clifford et al. (2015)	Loyalty	Loyalty	6.44	65%	4.29
A person gets a job in a foreign nation where their home country is viewed fairly negatively. In order to fit in among their coworkers, the person makes critical comments about their home country, though the person does not believe them.	Original	Loyalty	Loyalty	5.97	74%	4.15
A person cheats on their spouse with an ex-lover. The spouse does not find out about the infidelity.	Original	Loyalty	Loyalty	7.71	52%	7.45
Authority/Subversion Scenarios (Mean Wrongness: 5.20)						
A teenager takes their father's car out after curfew, against his wishes.	Clifford et al. (2015)	Authority	Authority	7.17	83%	5.50
An 18-year-old high school senior talks loudly in class while the instructor is teaching.	Original	Authority	Authority	7.82	93%	5.12
A staff member talks loudly and interrupts the mayor's speech to the public.	Clifford et al. (2015)	Authority	Authority	7.39	76%	5.00
A low-level company employee sends out an email to other low-level employees, calling the company president an "idiot."	Original	Authority	Authority	7.84	80%	4.73
A soldier disobeys an order given by their commanding officer. Nothing bad happens as a result of their disobedience.	Original	Authority	Authority	7.88	75%	4.88
An employee tries to undermine all of their boss's ideas in front of others.	Clifford et al. (2015)	Authority	Authority	7.06	58%	5.82
A person has a long and loud conversation with a group of friends during a religious service.	Clifford et al. (2015)	Authority	Authority	6.00	71%	5.35
Purity/Degradation Scenarios (Mean Wrongness: 5.24)						
A person eats the flesh of a deceased relative as part of a group funeral rite.	Clifford et al. (2015)	Purity	Purity	7.44	87%	6.52
A person searches through the trash to find women's discarded underwear.	Clifford et al. (2015)	Purity	Purity	7.75	87%	5.66
A person writes erotic poetry about their pet cat.	Original	Purity	Purity	6.56	84%	4.80
A person marries their first cousin in an elaborate wedding.	Clifford et al. (2015)	Purity	Purity	6.76	59%	4.57
A person takes part in a performance art piece in which a group of people act like animals for 30 minutes, including crawling around naked and urinating on stage.	Graham et al. (2009)	Purity	Purity	6.35	60%	3.58
A person looks at pornography in which an 18-year-old model has been digitally altered to look like she is 13.	Original	Purity	Purity	7.93	79%	6.34
A drunk elderly person offers to have oral sex with anyone in the bar.	Clifford et al. (2015)	Purity	Purity	6.69	74%	5.25
Miscellaneous Counter-Normative Scenarios (Mean Wrongness: 3.73)						
A person eats a T-bone steak with their hands in a fancy restaurant.	Huebner et al. (2010)	Counter-normative	Counter-normative	4.03	52%	2.81
A person tries to change the subject of conversation in the middle of their friend's	Original	Counter-normative	Counter-normative	4.75	53%	3.74

Scenario	Source	Hypothesized Foundation	Empirical Foundation	Foundation Score	%	Wrong
sentence.						
A person riding a bus plays music through their headphones so loudly that other passengers can hear it.	Huebner et al. (2010)	Counter-normative	Counter-normative	4.50	61%	4.03
A person walks away while their conversation partner is still speaking.	Original	Counter-normative	Counter-normative	4.93	52%	5.05
While in a rush, a person bumps into someone on the street, but does not say "excuse me."	Original	Counter-normative	Counter-normative	4.68	69%	3.42
A person goes to a birthday party for their friend, but does not bring him a gift.	Original	Counter-normative	Counter-normative	3.54	55%	3.03
A college student ignores their personal hygiene and mental health in order to devote more time to their studies.	Landy & Piazza (2017)	Non-moral	Counter-normative	4.91	50%	4.03
Non-Moral Scenarios (Mean Wrongness: 1.19)						
A person takes the bus to work instead of the train.	Original	Non-moral	Non-moral	1.33	100%	1.09
A person goes parasailing.	Original	Non-moral	Non-moral	2.12	95%	1.11
An accountant reads an advanced physics textbook for fun.	Original	Non-moral	Non-moral	1.47	100%	1.12
A person eats a blueberry muffin and drinks a glass of milk for breakfast.	Original	Non-moral	Non-moral	1.38	95%	1.13
A person wears their favorite outfit to a friend's party.	Original	Non-moral	Non-moral	1.30	97%	1.29
A person drives across town to visit their brother.	Original	Non-moral	Non-moral	1.47	98%	1.30
A person watches their favorite sports team on TV.	Original	Non-moral	Non-moral	1.24	95%	1.31

Note. "Empirical Foundation" is the foundation on which each action was rated most highly in Norming Study 2. If an action was rated as below the midpoint on all six foundations, it was classified as non-moral or counter-normative, depending on the results of Norming Study 3b. "Foundation Score" is the mean rating of how well a scenario exemplifies the foundation it was rated as most exemplifying (Norming Study 2). "%" is the percentage of participants who categorized a scenario as its empirical foundation in Norming Studies 3a and 3b. "Wrong" is the wrongness ratings from Norming Study 1.

Robustness Checks, Study 1

We re-ran the main analyses from Study 1 using other measures of conceptual relatedness. First, we used the minimum likelihood judgment in a pair of categories. For instance, if a participant rated the likelihood of committing a fairness violation, knowing that a person had committed a care violation, as 70%, and the likelihood of committing a care violation, knowing that a person had committed a fairness violation as 50%, that participant's fairness/care relatedness score would be 50%. Mean conceptual relatedness scores from this analysis are presented in Table S3. Next, we

used the maximum likelihood judgment within a pair of categories. So, for the participant who rated the likelihood of committing a fairness violation, knowing that a person had committed a care violation, as 70%, and the likelihood of committing a care violation, knowing that a person had committed a fairness violation as 50%, the relatedness score would be 70%. Mean conceptual relatedness scores from this analysis are presented in Table S4. Lastly, we used the product of likelihood judgments within a pair of categories. Thus, the relatedness score for this same participant in this analysis would be 70% x 50% = 35%. Mean conceptual relatedness scores from this analysis are presented in Table S5. Across all three of these alternative measures of relatedness, the results closely resembled those in the main text, with care, fairness, authority, and loyalty violations, and counter-normative actions, showing noticeably higher relatedness scores than loyalty and purity violations and non-moral actions.

Table S3. Mean conceptual relatedness scores, defined as the *minimum* likelihood judgment within category pairs.

	Care	**Fairness**	**Liberty**	**Authority**	**Loyalty**	**Purity**	**Non-Moral**	**Counter-Normative**
Care	60%	40%	32%	45%	39%	21%	23%	40%
Fairness		63%	33%	45%	41%	18%	29%	41%
Liberty			60%	28%	27%	15%	21%	28%
Authority				68%	42%	23%	25%	43%
Loyalty					61%	18%	23%	32%
Purity						36%	14%	21%
Non-Moral							54%	26%
Counter-Normative								57%
Mean Relatedness	**37%**	**38%**	**31%**	**40%**	**35%**	**21%**	**27%**	**36%**

Table S4. Mean conceptual relatedness scores, defined as the *maximum* of likelihood judgments within category pairs.

	Care	**Fairness**	**Liberty**	**Authority**	**Loyalty**	**Purity**	**Non-Moral**	**Counter-Normative**
Care	60%	64%	58%	68%	64%	54%	53%	67%
Fairness		63%	59%	69%	69%	51%	56%	65%
Liberty			60%	52%	51%	40%	53%	54%
Authority				68%	67%	54%	55%	70%
Loyalty					61%	52%	54%	59%
Purity						36%	50%	51%
Non-Moral							54%	56%
Counter-Normative								57%
Mean Relatedness	**61%**	**62%**	**53%**	**63%**	**60%**	**49%**	**54%**	**60%**

Table S5. Mean conceptual relatedness scores, defined as the *product* of likelihood judgments within category pairs.

	Care	Fairness	Liberty	Authority	Loyalty	Purity	Non-Moral	Counter-Normative
Care	44%	29%	22%	34%	29%	14%	14%	30%
Fairness		49%	23%	34%	31%	12%	18%	30%
Liberty			45%	18%	17%	10%	13%	18%
Authority				52%	32%	15%	16%	33%
Loyalty					45%	12%	14%	22%
Purity						36%	8%	14%
Non-Moral							36%	16%
Counter-Normative								41%
Mean Relatedness	**27%**	**28%**	**21%**	**29%**	**25%**	**15%**	**17%**	**26%**

As in the main text, we submitted the mean relatedness scores defined in these three ways to a hierarchical cluster analysis using between groups linkage. In agreement with our main analyses, violations of care, authority, fairness, and loyalty, and counter-normative actions were close to one another in Euclidean space and clustered together early in all three analyses. In contrast, violations of liberty and purity, and non-moral actions, were quite distant from the other categories. Figure S1 presents a dendrogram illustrating this analysis when relatedness scores are defined as minimum likelihood judgments, Figure S2 presents a dendrogram illustrating it when relatedness scores are defined as maximum likelihood judgments, and Figure S3 presents a dendrogram illustrating it when relatedness scores are defined as mean likelihood judgments.

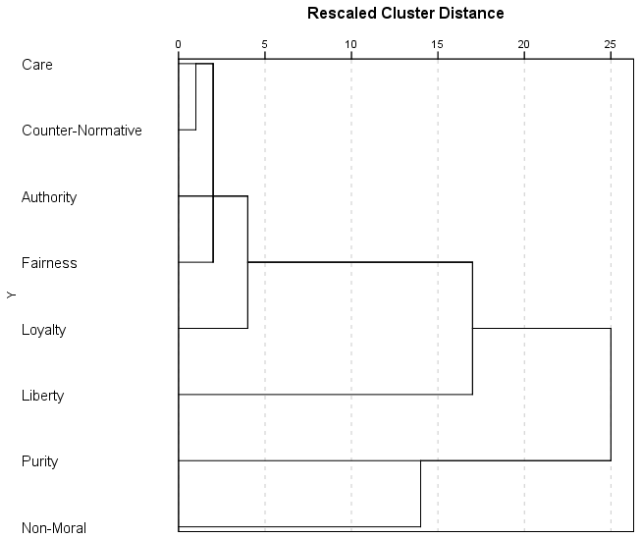

Figure S1. Dendrogram illustrating hierarchical cluster analysis (average linkage) of relatedness scores, defined as the minimum likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

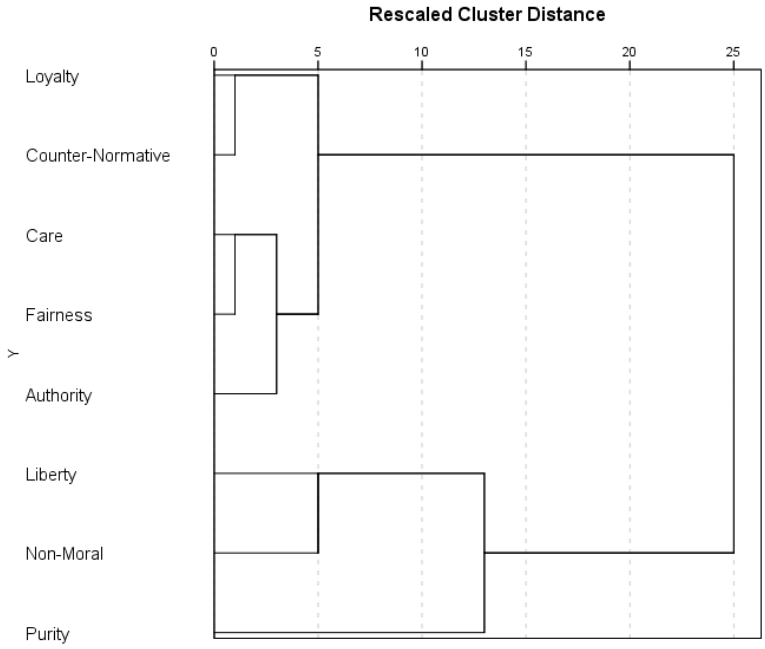

Figure S2. Dendrogram illustrating hierarchical cluster analysis (average linkage) of relatedness scores, defined as the maximum likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

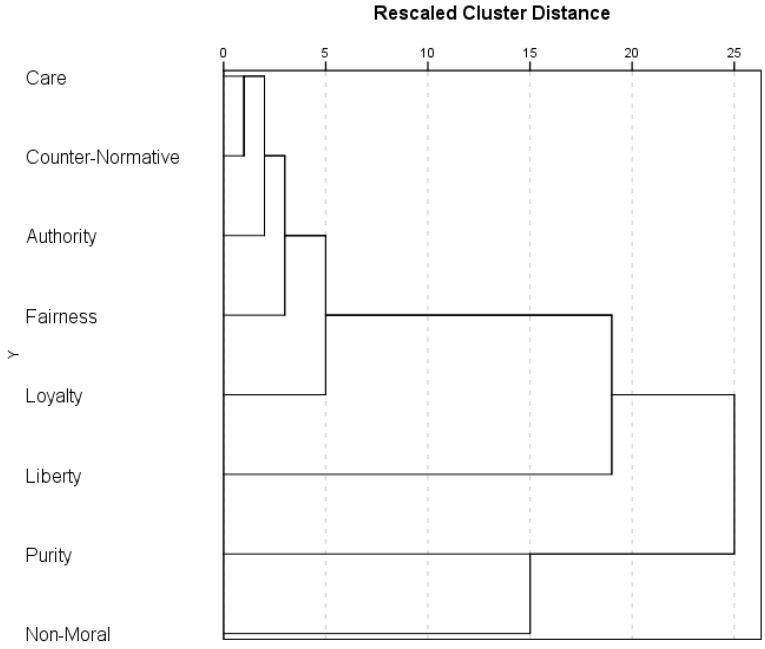

Figure S3. Dendrogram illustrating hierarchical cluster analysis (average linkage) of relatedness scores, defined as the product of likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

Next, we again subtracted the relatedness scores from 100% and submitted these dissimilarity scores to multi-dimensional scaling, treating the dissimilarity scores as ordinal variables. Figures S4, S5, and S6 present the two-dimensional solutions when relatedness scores are defined as the minimum, maximum, and mean likelihood judgment within a category pair, respectively. Once again, violations of care, authority, fairness, and loyalty, and counter-normative actions were quite close together in all three resultant two-dimensional spaces, while liberty violations, purity violations, and non-moral actions were more distant.

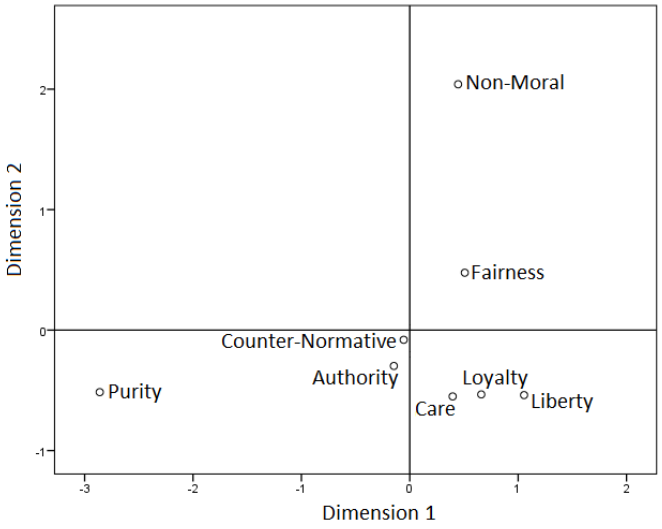

Figure S4. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as minimum likelihood judgments within a category pair. Model stress is .11.

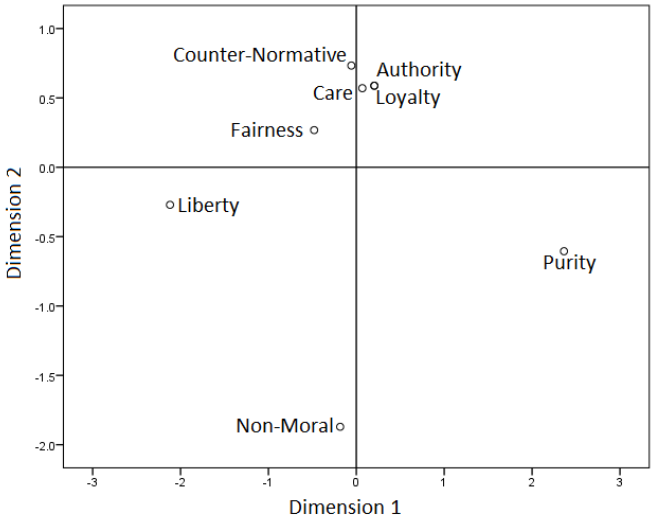

Figure S5. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as maximum likelihood judgments within a category pair. Model stress is .12.

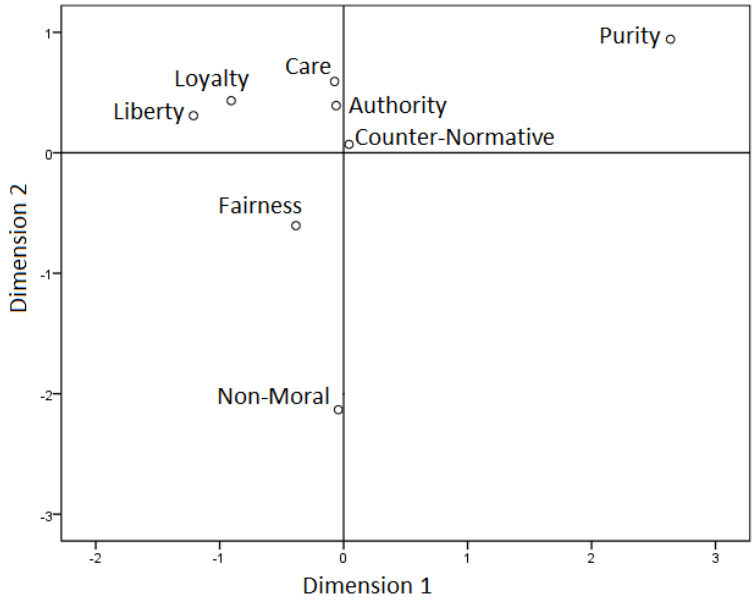

Figure S6. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as product of likelihood judgments within a category pair. Model stress is .11.

We also re-ran the cluster analyses and MDS analyses using all four definitions of conceptual relatedness (product, minimum, maximum, mean), excluding liberty violations, counter-normative actions, and non-moral actions, to test whether the results are consistent with our primary findings when examining only the five-foundation version of MFT. As can be seen below, no matter how the model is specified, care, fairness, authority, and loyalty violations were fairly close together in space, and purity violations were much more distant. This result agrees with our primary findings, and does not resemble the MFT-like taxonomy that would be expected based on the five-foundation version of MFT, which would categorize care and fairness under one superordinate category (individualizing foundations) and authority, loyalty, and purity under another (binding foundations). Interestingly, the dimensional structure uncovered in these analyses lends itself to a fairly clear interpretation, with Dimension 1 representing "not-purity" (or perhaps frequency or typicality). Dimension 2 might be interpreted as the inverse of harmfulness. This interpretation does not obviously apply to the analyses with all eight categories of action included, so we recommend taking it with a grain of salt.

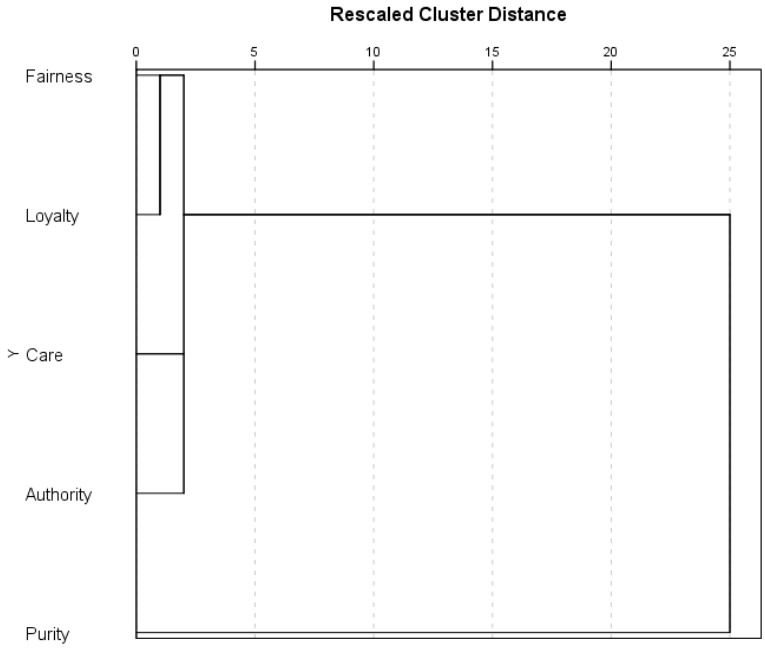

Figure S7. Dendrogram illustrating hierarchical cluster analysis of relatedness scores (average linkage), defined as the *mean* likelihood judgment within category pairs (liberty violations, counter-normative actions and non-moral actions excluded). *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

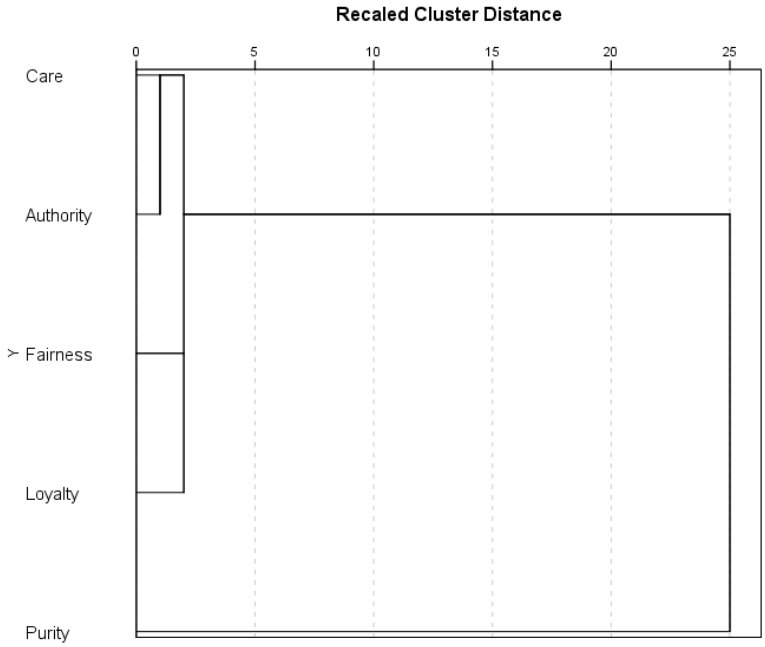

Figure S8. Dendrogram illustrating hierarchical cluster analysis of relatedness scores (average linkage), defined as the *minimum* likelihood judgment within category pairs (liberty violations, counter-normative actions and non-moral actions excluded). *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

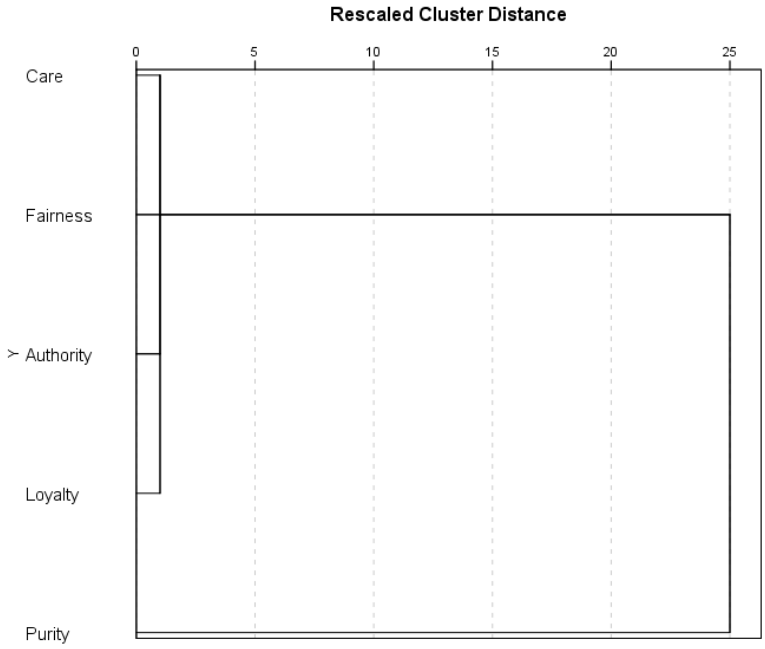

Figure S9. Dendrogram illustrating hierarchical cluster analysis of relatedness scores (average linkage), defined as the *maximum* likelihood judgment within category pairs (liberty violations, counter-normative actions and non-moral actions excluded). *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

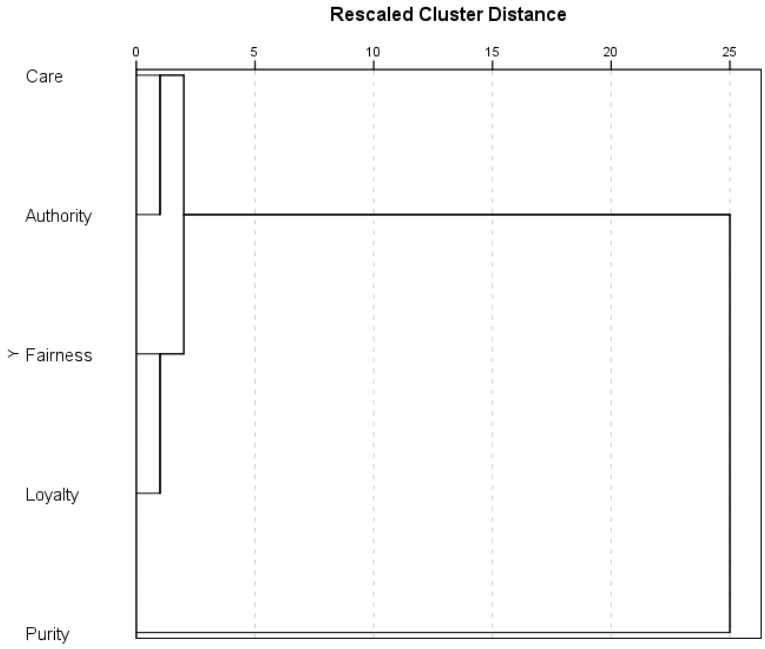

Figure S10. Dendrogram illustrating hierarchical cluster analysis of relatedness scores (average linkage), defined as the *product* of likelihood judgments within category pairs (liberty violations, counter-normative actions and non-moral actions excluded). *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

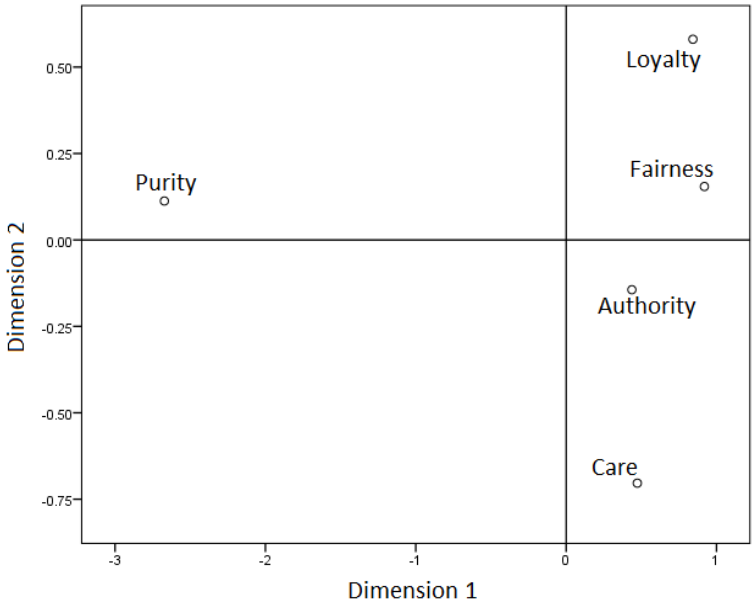

Figure S11. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as the *mean* likelihood judgment within a category pair. Model stress is .02.

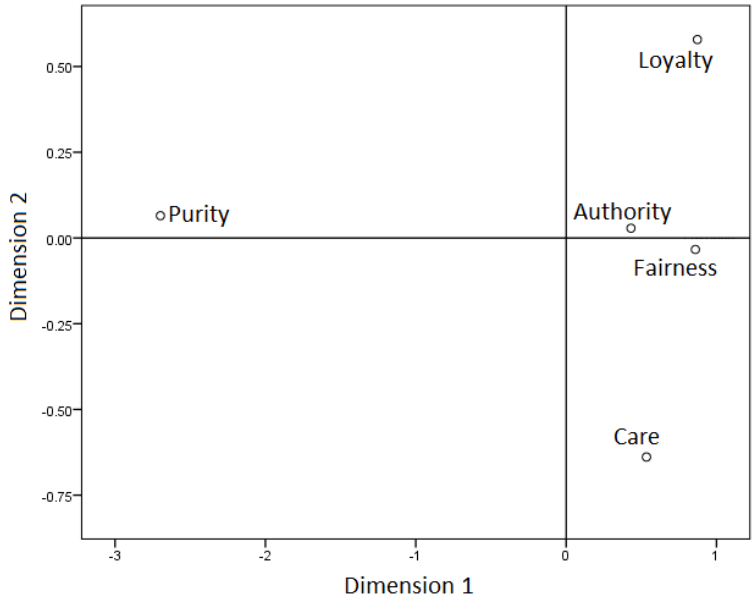

Figure S12. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as the *minimum* likelihood judgment within a category pair. Model stress is .01.

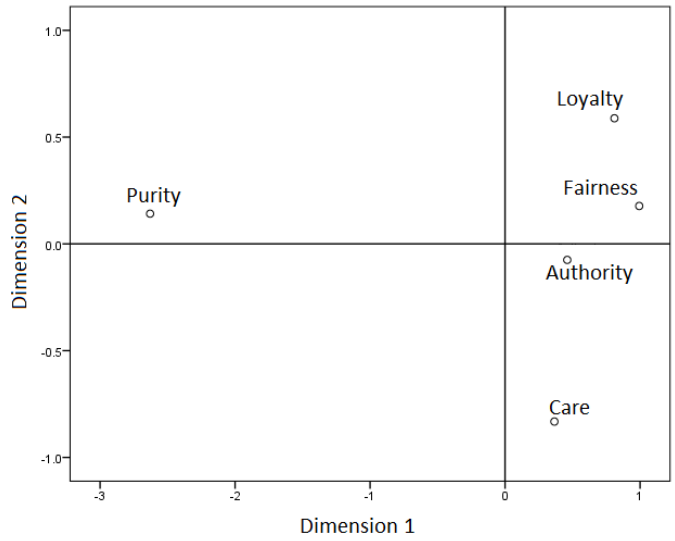

Figure S13. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as the *maximum* likelihood judgment within a category pair. Model stress is .01.

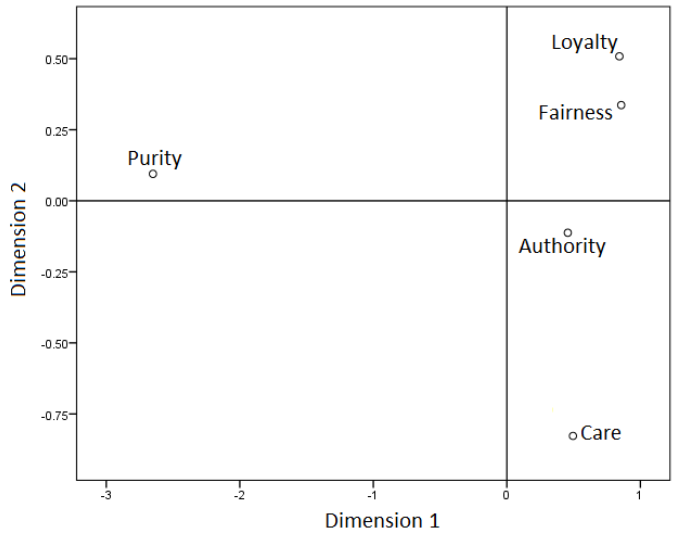

Figure S14. Two-dimensional solution derived from multi-dimensional scaling of relatedness scores, defined as the *product* of likelihood judgments within a category pair. Model stress is .07.

Lastly, we re-ran the hierarchical cluster analyses with all four types of conceptual relatedness scores, with single and complete linkage. As can be seen below, care, fairness, authority, and loyalty violations, and counter-normative actions, continue to cluster early, with liberty and purity violations and non-moral actions being far more distant.

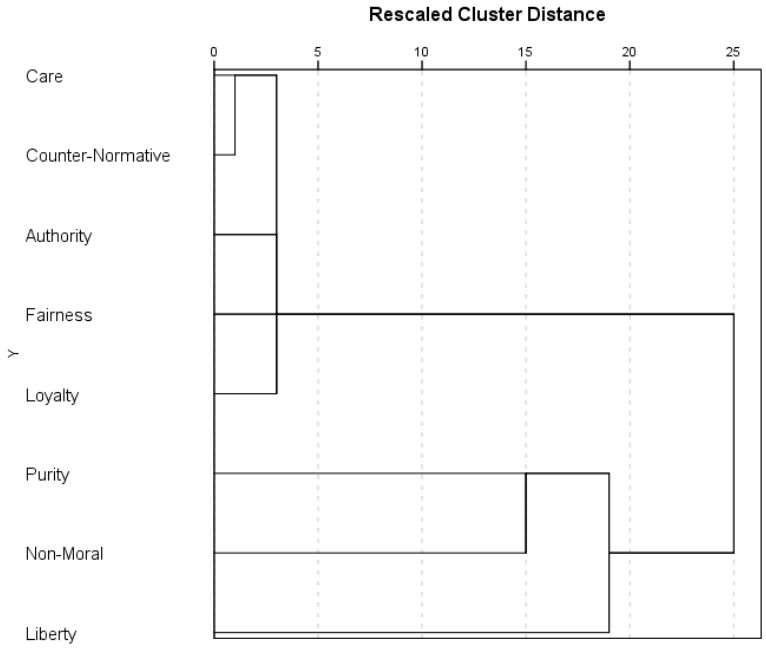

Figure S15. Dendrogram illustrating hierarchical cluster analysis (single linkage) of relatedness scores, defined as the *mean* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

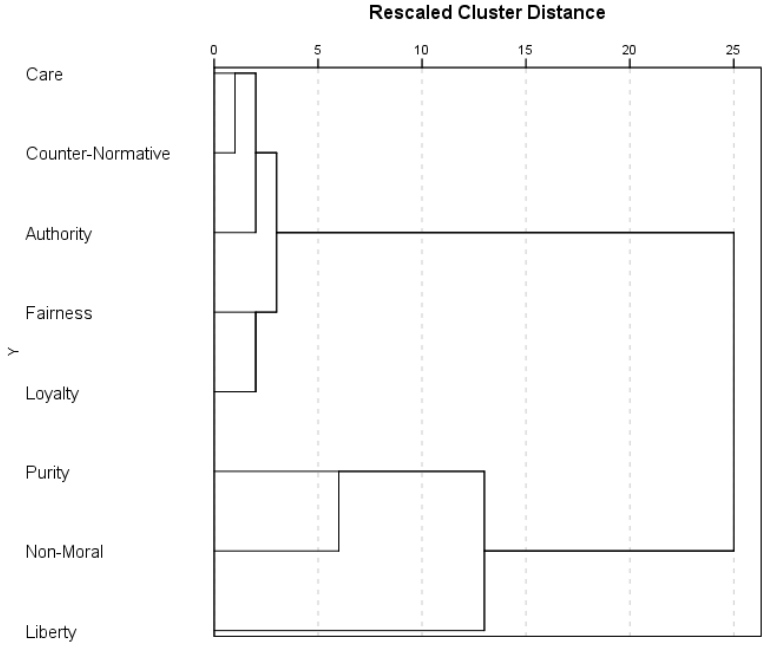

Figure S16. Dendrogram illustrating hierarchical cluster analysis (complete linkage) of relatedness scores, defined as the *mean* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

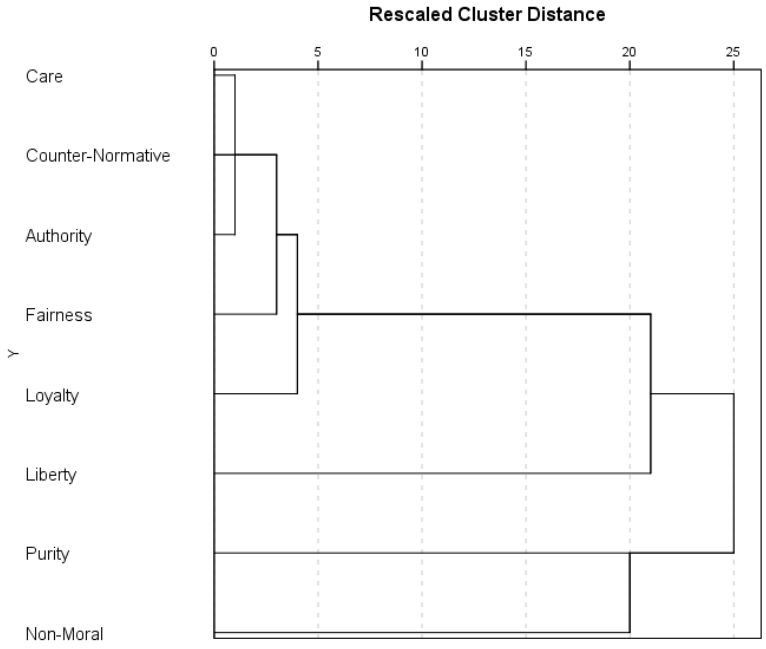

Figure S17. Dendrogram illustrating hierarchical cluster analysis (single linkage) of relatedness scores, defined as the *minimum* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

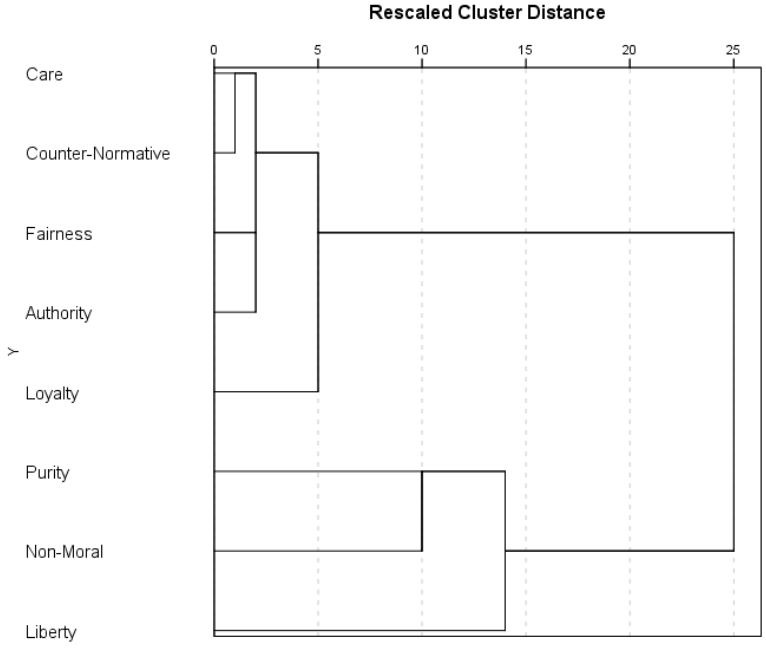

Figure S18. Dendrogram illustrating hierarchical cluster analysis (complete linkage) of relatedness scores, defined as the *minimum* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

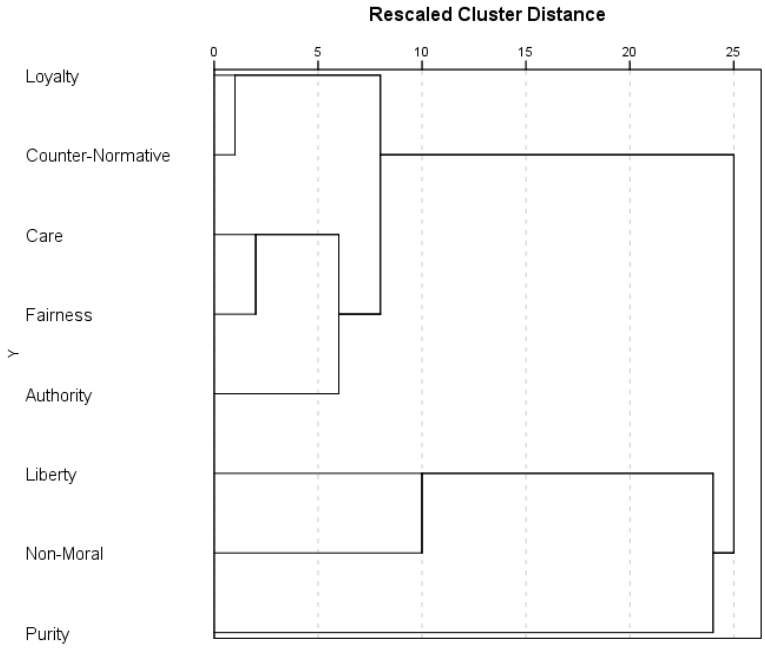

Figure S19. Dendrogram illustrating hierarchical cluster analysis (single linkage) of relatedness scores, defined as the *maximum* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

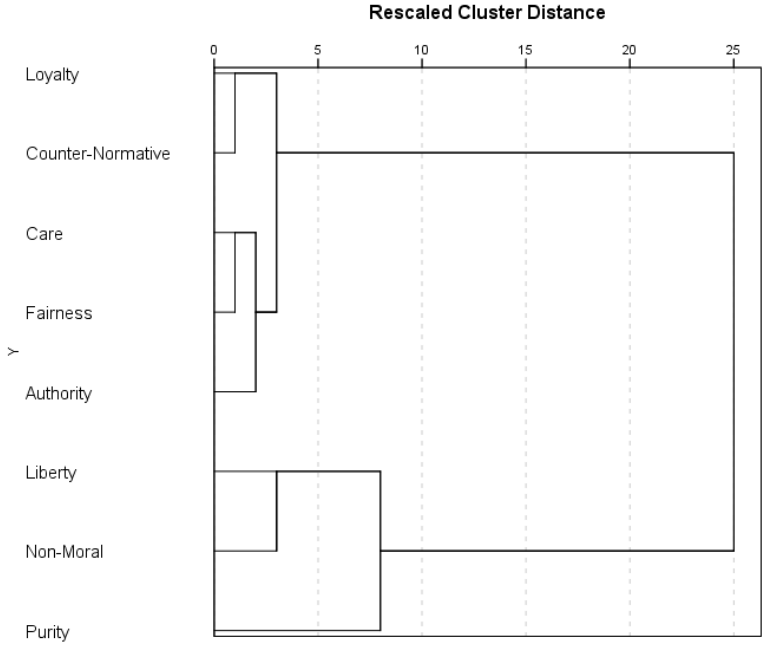

Figure S20. Dendrogram illustrating hierarchical cluster analysis (complete linkage) of relatedness scores, defined as the *maximum* likelihood judgment within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

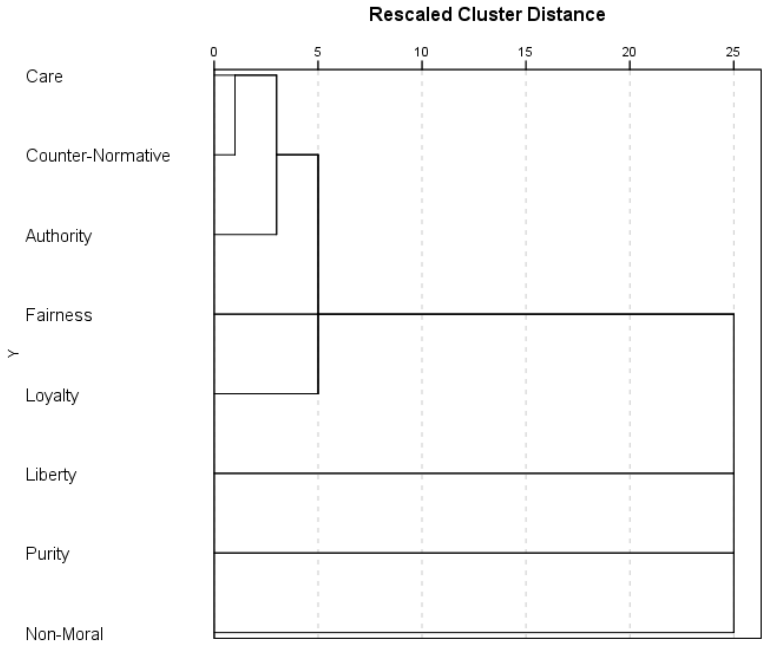

Figure S21. Dendrogram illustrating hierarchical cluster analysis (single linkage) of relatedness scores, defined as the *product* of likelihood judgments within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

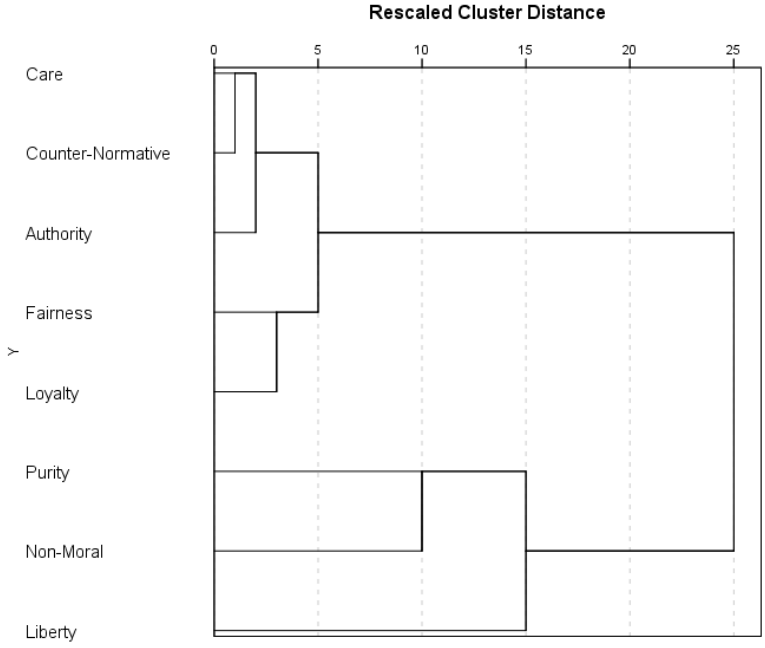

Figure S22. Dendrogram illustrating hierarchical cluster analysis (complete linkage) of relatedness scores, defined as the *product* of likelihood judgments within category pairs. *Note.* X-axis represents squared Euclidean distances between agglomerated clusters.

Linear Regressions, Study 1

To test whether participants' likelihood judgments merely reflect differences in the weirdness of different types of actions, we conducted linear regressions to statistically account for weirdness. The data set for these analyses is a "long form" of Study 1, in which each judgment is one line of data (64 lines per participant). We computed three models. In the first, we regressed likelihood judgments on the taxonomic distance between the premise and conclusion (a categorical variable coded as 0 for the same moral foundation, 1 for different foundations within the same superordinate category, and 2 for different superordinate categories). In the second, we regressed likelihood judgments on taxonomic distance, and on the mean weirdness ratings from Norming Study 1 of the premise and the conclusion that a participant judged. This model tests whether taxonomic distance still predicts likelihood judgments when statistically accounting for how weird the two actions being considered are. In the third model, we regressed likelihood judgments on taxonomic distance between the premise and the conclusion, and on the difference in weirdness between the premise and conclusion. Conceptually, an effect of this difference variable on likelihood judgments means that actions that are similar in how weird they are make each other more likely, i.e., a person who does one weird thing is seen as more likely to do another weird thing. In all three models, to account for the fact that each participant made multiple judgments, we clustered responses within participant and computed cluster-robust standard errors, using the lm.cluster command in the "miceadds" package for R (Robitzsch, Grund, & Henke, 2017). All three models are presented in Table S6. As can be seen, weirdness does significantly predict likelihood judgments. However, the predictive effect of taxonomic distance remains significant, and the coefficients on taxonomic distance do not change meaningfully when weirdness is included in the model. Thus, the structure of the derived taxonomy explains participants' inductive judgments over and above the weirdness of the actions being judged. We also conducted analogous analyses substituting frequency for weirdness. The results are essentially identical (see Table S7).

Table S6. Linear regressions with cluster-robust standard errors predicting likelihood judgments from taxonomic distance and weirdness (Study 1).

	Model 1	**Model 2**	**Model 3**
Intercept	59.26***	72.12***	59.41***
Taxonomic Distance = 1	-5.76***	-6.51***	-5.96***
Taxonomic Distance = 2	-20.96***	-20.78***	-21.08***
Premise Weirdness		1.75***	
Conclusion Weirdness		-4.69***	
Weirdness Difference			-3.21***
Observations		23,808	23,808
R^2	.08	.17	.16

Table S7. Linear regressions with cluster-robust standard errors predicting likelihood judgments from taxonomic distance and weirdness (Study 1).

	Model 1	**Model 2**	**Model 3**
Intercept	59.26***	38.32***	59.40***
Taxonomic Distance = 1	-5.76***	-6.42***	-5.96***
Taxonomic Distance = 2	-20.96***	-20.82***	-21.08***
Premise Weirdness		-1.57***	
Conclusion Weirdness		5.08***	
Weirdness Difference			3.33***
Observations		23,808	23,808
R^2	.08	.17	.15

Further Analyses, Studies 1-4

Table S.8. Correlations between participants' likelihood ratings (Studies 1 and 4) and confirmation measures (Study 3) and the taxonomies' predictions (1 = same category in derived taxonomy, different in MFT-like; -1 = same in MFT-like, different in derived; 0 = same or different in both taxonomies)

Study	**Median *r* (Derived vs. MFT)**	**% of Participants with *r* > .00**	***p* (Wilcoxon Test)**
1	.17	86.6	< .001
3	.04	61.6	< .001
4	.45	96.0	< .001

Table S.9. Correlations between participants' likelihood ratings and the taxonomies predictions, separately (1 = same superordinate category, 0 = different superordinate category)

Study	**Median *r* (Derived)**	**Median *r* (MFT-like)**	**% of Participants with higher *r* (Derived)**	***p* (Binomial Test)**
1	.32	.13	82.5	< .001
3	.19	.12	61.6	< .001
4	.45	-.45	96.0	< .001

Table S.10. Summary statistics for Anderson's *W*, an index of how well a participant's judgments fit the predictions of each taxonomy. $W > .50$ means the participant more closely resembles the predictions of the derived taxonomy.

Study	Median *W*	% of Participants with W > .50	*p* (Wilcoxon Test)
1	.55	82.5	< .001
3	.51	61.6	< .001
4	.62	94.1	< .001

Note. *W* is the parameter that minimizes squared error in the following mixture model. $SS = \sum [Y - Wx - (1-W)z]^2$, where *x* represents the predictions of the derived taxonomy and *z* represents the predictions of an MFT-like taxonomy (1 = same superordinate category, 0 = different superordinate category), and *Y* is the participant's score for each judgment (posterior probability judgments in Studies 1 and 4, and confirmation measures in Study 3; see Sanfey & Hastie, 2002). In our context, we compute *W* as: $W = \Sigma(x\text{-}z)(Y\text{-}z) / \Sigma(x\text{-}z)^2$.

Table S.11. Linear regression predicting Fisher-transformed within-subjects correlations (derived vs. MFT, see Table S.8) from demographic variables.

Variable	Study 1	Study 3	Study 4
Age	.00	.03	.01
Sex (Male)	.01	-.00	.04
Race/Ethnicity (Non-White)	-.04	-.02	-.20***
Politics (Conservatism)	.01	.08	.02
Education	.10†	.03	-.05

Table S.12. Linear regression predicting Anderson's *W* (Studies 1, 3, and 4) and average percentage of choices agreeing with derived taxonomy (Studies 2a and 2b) from demographic variables.

Variable	Study 1	Study 2a	Study 2b	Study 3	Study 4
Age	.04	.35**	.07	.03	.04
Sex (Male)	-.02	.02	-.09	-.02	-.07
Race/Ethnicity (Non-White)	-.02	.03	-.08	.00	-.12*
Politics (Conservatism)	-.02	.10	.06	.09	.02
Education	.06	-.09	.17	.04	-.01

Note: Average percentage of choices agreeing with derived taxonomy (Studies 2a and 2b) is equivalent to Anderson's *W* in these studies, excepting minor distortion due to the programming error in Study 2a [see Footnote 7 in main text])

Supplemental References

Chakroff, A. & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition, 136,* 30-37.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods, 47,* 1178-1198.

Graham, J. & Haidt, J. (2012). Sacred values and evil adversaries: A Moral Foundations approach. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* New York, NY: APA Books.

Haidt, J. (2012). *The righteous mind.* New York, NY: Vintage Books.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript,* University of Virginia.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or, is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613-628.

Huebner, B., Lee, J. J., & Hauser, M. D. (2010). The moral-conventional distinction in mature moral competence. *Journal of Cognition and Culture, 10,* 1-26.

Landy, J. F. & Piazza, J. (in press). Re-evaluating moral disgust: Sensitivity to many affective states predicts extremity in many evaluative judgments. *Social Psychological and Personality Science.*

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience, 23,* 3162-3180.

Piazza, J. & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making, 8,* 639-661.

Robitzsch, A., Grund, S., & Henke, T. (2017, August 24). Package 'miceadds'. Retrieved from https://cran.r-project.org/web/packages/miceadds/miceadds.pdf

Royzman, E. B., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the Divinity code. *Emotion, 14,* 892-907.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition, 112,* 159-174.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD Triad Hypothesis: A mapping between three moral emotions (contempt, anger, disgust), and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76,* 574-586.

Sanfey, A. G. & Hastie, R. (2002). Interevent relationships and judgment under uncertainty: Structure determines strategy. *Memory & Cognition, 30,* 921-933.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge, UK: Cambridge University Press.