Inductive Ethics: A Bottom-Up Taxonomy of the Moral Domain

Justin F. Landy (justinlandy@chicagobooth.edu)
University of Chicago
5807 S Woodlawn Avenue, Chicago, IL 60637 USA

Daniel M. Bartels (bartels@uchicago.edu)
University of Chicago
5807 S Woodlawn Avenue, Chicago, IL 60637 USA

Abstract

Moral Foundations Theory (MFT) posits that people moralize at least six distinct kinds of virtues. These virtues are divided into "individualizing" and "binding" virtues. Despite widespread enthusiasm for MFT, it is unknown how plausible it is as a model of people's conceptualizations of the moral domain. In this research, we take a bottom-up approach to characterizing people's conceptualization of the moral domain, and derive a taxonomy of morality that does not resemble MFT. We find that this model more accurately reflects people's theories of morality than does MFT.

Keywords: morality; inductive reasoning; concepts; categorization; taxonomies

Introduction

How do people conceptualize the structure of the moral domain? Despite the recent explosion of research in the cognitive science of morality, a satisfactory empirical answer to this foundational question has not yet emerged.

An early attempt to understand conceptualizations of morality focused on the distinction between acts that violate moral principles, and acts that violate social conventions (Turiel, 1983). "Domain Theory" (DT) posits that moral violations concern "justice, rights, and welfare" (Turiel, 1983, p. 3), and that other prohibited actions, though condemnable, only constitute violations of convention, and could be permissible under alternative normative systems.

However, more recent evidence suggests that at least some people treat some acts that cause no objective harm and violate no rights as being truly morally wrong; people's moral domains are more complex than DT would suggest (Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993; Landy, 2016; Royzman, Landy, & Goodwin, 2014). The most prominent model of this complexity is Moral Foundations Theory (MFT; Graham, Haidt, & Nosek, 2009; Haidt & Joseph, 2004). In brief, MFT posits that there are at least six "moral foundations" – distinct, sometimes competing, virtues that we are innately prepared to moralize. These six virtues are divided into "individualizing" foundations that are concerned with the rights and welfare of individuals – harm prevention ("care"), fairness, and liberty – and "binding" foundations that are concerned with preserving the moral community – loyalty, respect for and obedience to authority, and bodily and spiritual purity ("sanctity"). Figure 1 presents the taxonomic structure of MFT.

MFT was developed by joining insights from cultural psychology and anthropology with evolutionary reasoning, and it has not been tested as a model of people's moral concepts. Thus, despite the theory's popularity, it remains unknown whether it is a plausible model of people's cognitive representations of the moral domain. In this research, we test the plausibility of MFT as such a model, using methods from the study of inductive reasoning. Specifically, we use people's inductive judgments about likely behaviors to derive a bottom-up model of people's taxonomy of the moral domain, and compare this with the theoretical taxonomy proposed in MFT.

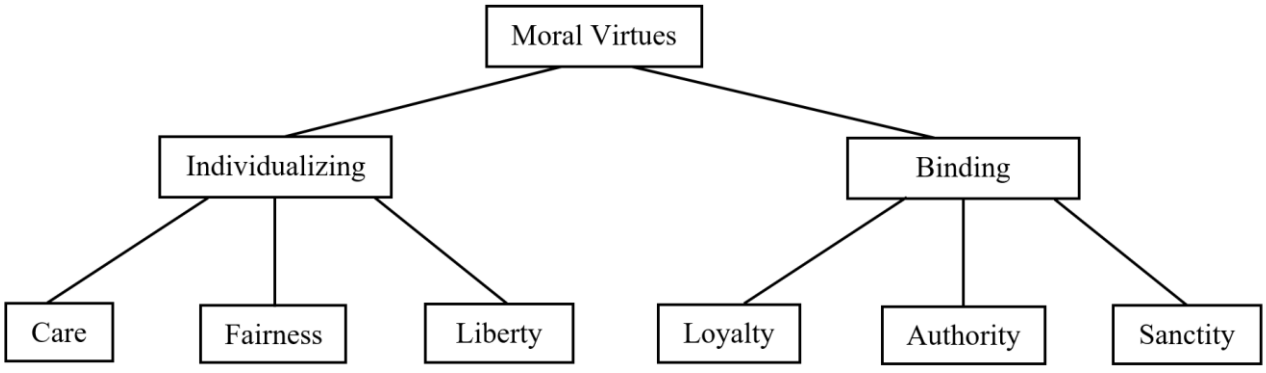

Figure 1: Moral Foundations Theory.

Category-Based Induction and Moral Virtues

The study of category-based induction (Osherson, Smith, Wilkie, López, & Shafir, 1990) is based on the premise that concepts are represented 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 is indicative of how closely related the objects are in that person's taxonomy. An example, adapted from Osherson et al. (1990), illustrates this: consider the premise "Robins use serotonin as a neurotransmitter." Among the (infinite) possible conclusions, "sparrows use serotonin as a neurotransmitter" is typically 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. More formally, there is greater premise-conclusion similarity in the former case than in the latter. Robins and sparrows might, for instance, both belong to the superordinate category "songbirds", whereas geese would belong to a separate superordinate category.

We applied this same logic to taxonomies of moral virtues, with the specific aim of testing the validity of MFT as a model of these taxonomies. Consider the premise "Joe commits a fairness violation." If MFT's taxonomy is a reasonable model of people's conceptualizations of the moral domain, then the conclusion "Joe would also commit a care violation" should be considered more likely to be true than "Joe would also commit a loyalty violation." Et cetera.

In Study 1, we had participants rate the likelihood that a person would engage in a wide variety of actions that exemplify the six moral foundations (conclusions), given information about a previous behavior (premise). From these likelihood ratings, we extracted a bottom-up taxonomy of the moral domain. In Study 2, we used a similar task, but also included ratings of the baseline likelihood of each behavior (i.e., the likelihood of a conclusion in the absence of any premise). The increases in inductive strength gained from the inclusion of premises more closely resembled the predictions of our taxonomy than MFT. Finally, in Study 3, we presented participants with two premises, rather than one. Participants' likelihood judgments more closely resembled the predictions of our taxonomy than MFT.

Stimulus Development

Stimuli in all three studies consisted of the long form of the Moral Violations Database – Severity Equated (MVD-SE), a subset of the Moral Violations Database (MVD), a set of nearly 250 behavioral descriptions normed on several criteria, including moral wrongness and representativeness of each moral foundation. Most of the stimuli included in the MVD are original, or are modified forms of the Moral Foundations Vignettes (Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015). The development of the MVD and MVD-SE is detailed elsewhere (Landy & Bartels, 2016), so we only briefly summarize it here.

The MVD-SE contains seven behaviors violating each foundation, drawn from the larger MVD. These behavioral descriptions passed a two-step validation process: one sample rated how well each behavior exemplified each moral foundation, then a second sample assigned each behavior to the foundation that it best exemplified, in a forced-choice task. Behaviors that were rated above the scale midpoint for a foundation by the first sample, and assigned to that foundation by a majority of the second, were considered validated. From these validated stimuli, seven were chosen to represent each foundation (e.g., a person drives past a man on an empty road who is clearly injured (care), hires their nephew instead of a more qualified job applicant (fairness), forces their daughter to enroll as a pre-med student in college (liberty), sends out an email calling their boss an "idiot" (authority), makes critical comments about their home country (loyalty), or looks at pornography in which an adult model has been digitally altered to look like she is 13 years old (sanctity)). 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). Moreover, the mean moral wrongness ratings for the foundations are extremely closely equated (5.20-5.26, on a 1-9 scale). The MVD-SE also includes seven non-moral actions, which extensive pretesting has found to be morally inert (e.g., "a person goes parasailing"), and seven counter-normative behaviors 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 counter-normative actions largely consist of violations of polite etiquette.

Study 1

Method

Participants (N = 367) were recruited online through Amazon Mechanical Turk in exchange for monetary compensation. The study was completed online.

Each participant made 64 likelihood judgments, one for each possible premise/conclusion combination of the eight conceptual categories in the MVD-SE (e.g., authority/authority, authority/non-moral, etc.). Premises and conclusions were randomly sampled from the MVD-SE for each question, with the restriction that the premise and conclusion could not be the same action. Questions took the following form: "A person hires their nephew for a job, instead of a more qualified applicant. 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 x 7 conclusions = 49 possible

[1] There are, of course, numerous other models of how concepts are represented (see Medin, Rips, & Smith, 2005, for a review). We follow the majority of the research on category-based induction, which we see our methodological approach as deriving from, in focusing on taxonomic representations.

Table 1: Mean conceptual relatedness scores.

	Care	Fairness	Liberty	Authority	Loyalty	Sanctity	Non-Moral	Counter-normative
Care	44%	25%	22%	34%	29%	14%	14%	30%
Fairness		49%	23%	34%	31%	11%	18%	30%
Liberty			45%	18%	17%	10%	13%	18%
Authority				52%	32%	15%	16%	33%
Loyalty					45%	12%	14%	22%
Sanctity						36%	8%	14%
Non-Moral							36%	16%
Counter-Normative								41%

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").

Results

For present purposes, we computed a measure of conceptual relatedness between categories of actions by multiplying likelihood estimates to and from pairs 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% x 50% = 35%.

We submitted the mean relatedness scores (presented in Table 1) to a hierarchical cluster analysis using between groups linkage.[2] In agreement with the pattern of means in Table 1, violations of care, authority, fairness, and loyalty, and counter-normative actions were close to one another in Euclidean space and clustered together early in the analysis. In contrast, violations of liberty and sanctity, and non-moral actions were quite distant from all other categories. Figure 2 presents a dendrogram illustrating this analysis.

We confirmed this result by subtracting relatedness scores from 100%, and submitting the resulting dissimilarity scores to multi-dimensional scaling.[3] We restricted our analysis to a two-dimensional solution for ease of presentation, and treated the dissimilarity scores as ordinal variables.[4] As shown in Figure 3, and consistent with the above analyses, violations of care, authority, fairness, and loyalty, and counter-normative actions are quite close to one another in the resultant two-dimensional space, with liberty violations, and especially sanctity violations and non-moral actions, more distant. Model stress was .10, which is generally considered acceptable (see, e.g., Kruskal, 1964a, 1964b; Rosenberg, Nelson, & Vivekananthan, 1968).

These analyses converge on the conclusion that care, fairness, authority, and loyalty violations, and counter-normative actions, are quite closely related to one another in people's taxonomies of morality, while liberty and sanctity violations and non-moral actions are less closely related. It seems reasonable, therefore, to model the virtues of care, fairness, authority, and loyalty, along with politeness, as belonging to a single superordinate category. We conceptualize this category as "obedience to rules". We think that this captures what sets these virtues apart from liberty and sanctity – liberty has to do with not *creating* rules that are burdensome or oppressive for others, and

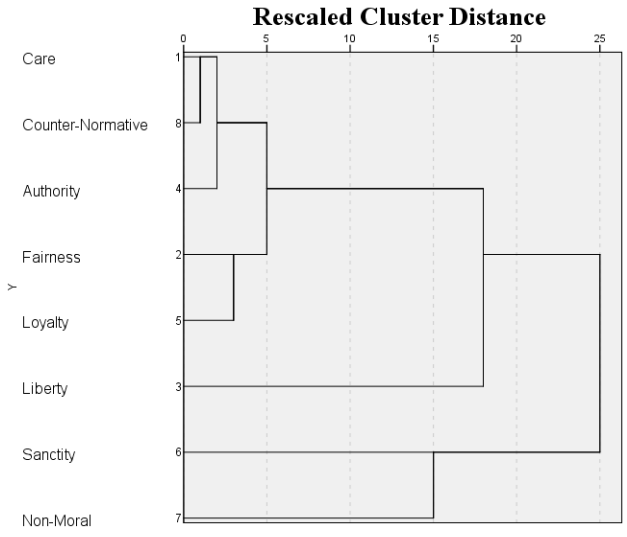

Figure 2: Dendrogram illustrating hierarchical cluster analysis of relatedness scores. X-axis represents squared Euclidean distances between agglomerated clusters.

[2] The results are essentially identical when Ward's method is used instead. We present the results of the analysis using between groups linkage because the resulting dendrogram makes the relationships between categories easier to visualize.

[3] Identical results are obtained if the dissimilarity scores are calculated by subtracting relatedness scores from the maximum observed relatedness (52%) than from the maximum possible relatedness (100%). We therefore focus on the conceptually simpler analysis.

[4] The pattern of results is the same – indeed, it is somewhat clearer – if the dissimilarity scores are treated as interval or ratio variables, however, the model stress is unacceptably high under these assumptions (.20 and .35, respectively).

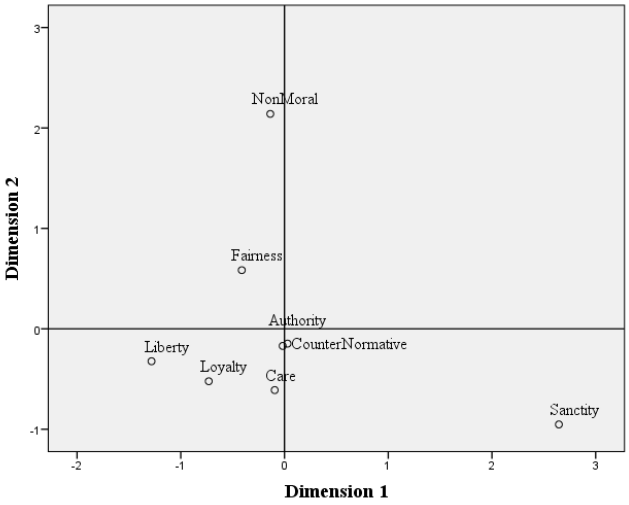

Figure 3: Two-dimensional solution derived from multi-dimensional scaling of relatedness scores.

sanctity violations tend to be so unusual that explicit rules forbidding them (e.g., "Thou shalt not write erotic poetry about thy cat") are probably rarely articulated. Our bottom-up taxonomy of the moral domain is presented in Figure 4. Note that this taxonomy does not resemble MFT, and that the individualizing-binding distinction did not emerge in our analyses of people's judgments.

Study 2

Method

Participants (N = 359) were recruited in the same manner as in Study 1. Participants from Study 1 could not take part in this study.

Each participant made 42 total likelihood judgments, one for each possible premise/conclusion combination of the six moral foundations, and six baseline likelihood judgments with no premise. Politeness and non-moral characteristics are not included in MFT's taxonomy; therefore we did not include the counter-normative and non-moral actions from Study 1 in Studies 2 and 3, as they are not useful for testing the relative predictive validities of the two taxonomies. 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. That is, each participant saw the same conclusion from every foundation seven times, so that their premised judgments were directly comparable to their baseline likelihood judgments. Likelihood ratings were made on the same sliding scale as in Study 1.

Results

We created a measure of inductive strength gained from knowledge of a premise by subtracting baseline likelihood judgments from premised judgments (e.g., if a participant rated the baseline likelihood of a person hiring their nephew instead of a more qualified applicant at 30%, and the likelihood of this, given that the person had driven past an injured man on an empty road without stopping to help, at 65%, the gain in inductive strength would be 35%).

Both taxonomies classify 18 premise-conclusion pairs as belonging to the same superordinate category (e.g., authority and care are both part of obedience to rules in our taxonomy, and loyalty and sanctity are both binding virtues in MFT), and 18 as belonging to different superordinate categories (e.g., liberty and sanctity in both taxonomies). Thus, we calculated the average inductive strength gained from premises that belong to the same superordinate category as the conclusion, versus premises that do not, in each taxonomy. We expected a Premise x Taxonomy interaction, such that more inductive strength would be gained from within-category premises, versus between-category premises, according to our taxonomy, versus MFT.

Means and standard deviations are presented in Table 2. Substantially more inductive strength was gained from within-category premises than between-category premises in both taxonomies (within-subjects ANOVA: $F(1, 358) = 469.46$, $p < .001$, $\eta^2_p = .57$), and there was no mean difference in inductive strength gained across taxonomies ($F(1, 358) = .036$, $p = .850$). However, as expected, the

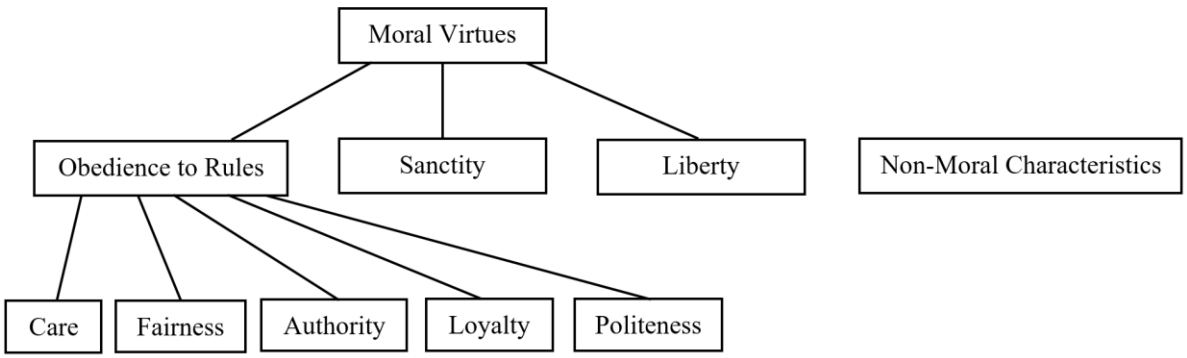

Figure 4: A bottom-up taxonomy of the moral domain.

Table 2: Mean gain in inductive strength from within-category premises and between-category premises.

	MFT	Bottom-Up
Within-Category	11.63 (12.07)	13.30 (13.65)
Between-Category	5.27 (11.29)	3.64 (11.51)

difference in strength gained from within-category and between-category premises was significantly larger for our taxonomy than for MFT, again suggesting that our taxonomy is a better model of conceptualizations of morality than MFT (interaction $F(1, 358) = 21.92$, $p < .001$, $\eta^2_p = .058$).

We next examined cases in which the two taxonomies make differing predictions regarding what premise should be more informative about a given conclusion. For example, in our taxonomy, fairness belongs to the same superordinate category as authority, while sanctity does not; therefore, our taxonomy predicts that the gain in inductive strength for authority conclusions will be greater when participants are given fairness premises than when they are given sanctity premises. In MFT, the reverse is true; MFT therefore makes exactly the opposite prediction. There are eight such combinations, presented in Table 3. Paired- sample t-tests of the gains in inductive strength generally agreed with the predictions of our taxonomy – five of eight significantly supported it, and none significantly supported MFT.

Finally, we constructed a matrix comparing the categorizations in our taxonomy with those in MFT. Premise-conclusion pairs which are in the same superordinate category in our taxonomy but not in MFT (e.g., care/authority) were coded as 1, pairs which are in the same superordinate category in MFT but not in our taxonomy (e.g., care/liberty) were coded as -1, and pairs that both taxonomies categorize in the same way were coded as 0. We computed a correlation between this matrix of categorizations and expressed gains in inductive strength for each participant. A positive correlation indicates that a participant's judgments conform more to the predictions of our derived taxonomy than to those of MFT, and a negative correlation indicates the opposite.

Table 3: Within-Subjects *t*-tests of Study 2 Predictions. *Note:* *$p < .05$; ***$p < .001$.

	Within-Category Premises		
Conclusion	**MFT**	**Bottom-Up**	***t*(358)**
Care	Liberty	Authority	3.99***
Care	Liberty	Loyalty	2.17*
Fairness	Liberty	Authority	4.18***
Fairness	Liberty	Loyalty	3.62***
Authority	Sanctity	Care	4.90***
Authority	Sanctity	Fairness	.40, *ns*
Loyalty	Sanctity	Care	-.008, *ns*
Loyalty	Sanctity	Fairness	-.88, *ns*

Two-hundred-twenty participants out of 358 (61%)[5] expressed a positive correlation, which a binomial test suggests is unlikely to be due to chance, $p < .001$. Moreover, the median correlation, $r = .042$, is significantly larger than 0, by a one-sample Wilcoxon signed-rank test, $p < .001$. Finally, a one-sample t-test performed on the Fisher-transformed correlations indicates that the mean, $z = .047$, is significantly greater than zero, $t(357) = 5.15$, $p < .001$, $d = .27$. Thus, participants' judgments conformed more to the predictions of the taxonomy derived in Study 1 than to the predictions of MFT.

Thus, regardless of how they are analyzed, the gains in inductive strength from learning about a prior behavior consistently resembled the predictions of our taxonomy more than those of MFT. This provides confirmatory evidence that our derived framework better describes people's conceptualizations of morality than does MFT.

Study 3

Method

Participants ($N = 363$) were recruited in the same manner as in the previous studies. Participants from Studies 1 and 2 could not take part in this study.

There are six combinations of two premises and a conclusion in which the two taxonomies make differing predictions about whether the conclusion belongs to the same superordinate category as the premises (see Table 4).

Participants made 24 likelihood judgments of the same form as in Studies 1 and 2, but with two premises instead of one. The premise-conclusion combinations were randomly selected, with the restriction that each participant received four instances of each of the six combinations for which the two taxonomies make differing predictions.

Results

As in Study 2, we created a matrix of categorizations derived from the two taxonomies. Premise-conclusion combinations that belong to the same superordinate category in our taxonomy, but not MFT, were coded as 1, whereas combinations that belong to the same superordinate category in MFT, but not our taxonomy, were coded as -1.

Table 4: Study 3 Predictions.

Premises	**Conclusion**	**Taxonomy Predicting Same Category**
Care/Fairness	Liberty	MFT
Care/Fairness	Authority	Bottom-Up
Care/Fairness	Loyalty	Bottom-Up
Authority/Loyalty	Sanctity	MFT
Authority/Loyalty	Care	Bottom-Up
Authority/Loyalty	Fairness	Bottom-Up

[5] One participant responded "50%" to every question, expressing no variance in her judgments. Her data were excluded from these analyses.

We then computed within-subjects correlations between these codes and participants' likelihood judgments, as in Study 2. A positive correlation indicates that a participant's judgments conform more to our predictions than to MFT, while a negative correlation indicates the opposite. Three hundred-forty-eight participants (96%) expressed a positive correlation, while only 15 (4%) expressed a negative or zero correlation, a result which is highly unlikely to be due to chance, binomial test $p < .001$. Moreover, the median correlation, $r = .45$, is significantly greater than zero, by a one-sample Wilcoxon signed-rank test, $p < .001$. Finally, the mean Fisher-transformed correlation, $z = .48$, is significantly greater than zero, $t(362) = 35.13$, $p < .001$, $d = 1.84$. As in Study 2, participants' judgments conformed more to the predictions of our derived taxonomy than to those of MFT.

General Discussion

Three studies converged on the conclusion that Moral Foundations Theory (MFT) does not describe people's taxonomies of moral virtues well. In Study 1, we derived a taxonomy of virtues, which does not resemble MFT. Study 2 examined the increases in inference strength that come from knowledge of prior behavior, and found that the strength of these gains aligned more closely with the predictions of our taxonomy than those of MFT. Lastly, Study 3 found that ratings of the likelihood of behaviors, given information about two prior behaviors, conformed more to the predictions of our taxonomy than MFT. It is important to note, however, that our results speak *only* to the plausibility of MFT as a model of people's theories of morality, and do not bear on its validity as an evolutionary model of variation in moral virtues.

We do not claim that our derived taxonomy represents the most comprehensive model possible of people's theories of morality. Indeed, because we used stimuli that were already known to be uniquely good exemplars of the moral foundations, we may have left out elements of the moral domain that MFT overlooks. We think that a particularly strong candidate for such an overlooked virtue is honesty, which is sometimes considered part of the fairness foundation, but seems intuitively to be valued even in the absence of fairness concerns (see Landy & Uhlmann, 2016, for a discussion of honesty in folk virtue ethics). Developing a fully bottom-up model of the moral domain that does not inherit the assumptions of any theory is a difficult task, but such a model would be very informative and could help to develop new theoretical advances. Therefore, we see the development of a more comprehensive mapping of concepts of morality as an important direction for future research.

In conclusion, MFT does not seem to model people's conceptualizations of the moral domain especially well. In particular, the distinction between individualizing and binding virtues does not seem to reflect a psychologically real division that people make. By providing a more accurate picture of how people parse their moral worlds, this research helps to clarify a fundamental question in the cognitive science of morality.

Acknowledgments

We thank Halley Bayer for assistance in conducting this research, and Stephanie Chen and the members and guests of the Morality Research Lab for valuable feedback.

References

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

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.

Haidt, J. & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31,* 191-221.

Haidt, J. & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133,* 55-66.

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.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: I. *Psychometrika, 29,* 1-27.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: II. *Psychometrika, 29,* 115-129.

Landy, J. F. (2016). *Form and content in the categorization of moral violations.* Chicago, IL: University of Chicago.

Landy, J. F. & Bartels, D. M. (2016). *The Moral Violations Database.* Chicago, IL: University of Chicago.

Landy, J. F. & Uhlmann, E. L. (2016). *Morality is personal*. Forthcoming in K. Gray & J. Graham (Eds.), *The atlas of moral psychology.* New York: Guilford.

Medin, D. L. & Rips, L. J. (2005). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 37-72). Cambridge, UK: Cambridge University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185-200.

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.

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.

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