Social Psychological and
Personality Science
1-4
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550615592243
spps.sagepub.com
SAGE

# Disconfirming Moral Foundations Theory on Its Own Terms: Reply to Graham (2015)

Kurt Gray[1] and Jonathan E. Keeney[2]

## Abstract

People see immorality in sin and sex, but is "purity" a unique type of moral content, with unique cognition? Domain-general accounts—and parsimony—suggest that all moral content is processed similarly and that "purity" is merely a descriptive label. Conversely, domain-specific theories (e.g., moral foundations theory [MFT]) argue for a special purity module. Consistent with domain-general accounts, we demonstrated that purity concerns are not distinguished from harm concerns—in either MFT or naturalistic scenarios—and that controlling for domain-general dimensions eliminates effects previously ascribed to moral "modules." Here, we reaffirm the strength of our data, exploring how issues raised by Graham reflect only weaknesses in MFT. Importantly, we identify several clear contradictions between Graham's comment and past-published accounts of MFT. To the extent that MFT stands by its published stimuli, methodologies, and theoretical assumptions, we believe that we have disconfirmed MFT on its own terms.

## Keywords

justice, morality, social cognition, values, ethics/morality

Biological science has revealed staggering diversity among organisms—biological pluralism—but all biological species derive from the same general process of evolution. Likewise, anthropology has revealed staggering diversity among moral judgments—moral pluralism—but all moral judgments may revolve around the same domain-general, harm-based dyadic moral template (Gray & Schein, 2012). Even the founder of moral pluralism, Richard Shweder (2012), advocates for "universality without uniformity" in which violations of "purity" can be understood via perceived harm.

Arguing against a common moral template, moral foundations theory (MFT) suggests that moral judgment occurs via discrete cognitive modules: "little switches in the brain of all animals" "triggered" by "specific moral inputs," such as harm or purity (Haidt, 2012, p. 123), with "distinct cognitive computations" for each kind of moral content (Young & Saxe, 2011, p. 203). The main evidence for these claims comes from researcher-constructed scenarios of "harm" (e.g., murder) and "purity" (e.g., chicken masturbation) that reveal different patterns of judgment (Graham et al., 2013). However, our research finds that these scenarios fundamentally confound moral content (harm vs. purity) with domain-general dimensions, including severity and weirdness (among potential others; Gray & Wegner, 2011). In our controlled studies, purity per se demonstrates no special effect on moral cognition, nor does it appear to be distinct from harm—or even pass manipulation checks—all arguing against MFT modularity (Gray & Keeney, 2015).

In his comment, Graham (2015) criticized four elements of our research, namely, (1) our use of moral judgment scenarios as stimuli; (2) our reliance upon participant intuitions in categorizing moral content; (3) the high correlations between harm, purity, and severity; and (4) our description of MFT as domain specific. Although these criticisms seem superficially compelling, a closer examination reveals that they actually highlight weaknesses within MFT. As our stimuli, design choices, analysis logic, and theoretical descriptions all come directly from MFT, any flaws therein argue only against MFT. Even more problematic, Graham's comment redefines MFT in ways that seem impossible to reconcile with past published accounts. Through our specific points, we elaborate on these inconsistencies and explain how they undermine the coherence of MFT.

## *Moral Judgment Scenarios*

Graham criticizes our research for operationalizing purity primarily though MFT moral judgment scenarios, rather than other MFT stimuli. Although there may be other MFT items,

[1] Department of Psychology, University of North Carolina, Chapel Hill, NC, USA
[2] Kenan-Flagler Business School, University of North Carolina, Chapel Hill, NC, USA

**Corresponding Author:**
Kurt Gray, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599, USA.
Email: kurtgray@unc.edu

that is beside the point—shouldn't all MFT stimuli be free of sampling bias? More importantly, we specifically chose these scenarios in order to test MFT on its own terms. Graham and colleagues (2009) custom developed these scenarios to represent purity and frequently cite them as providing strong support for MFT (Graham et al., 2013). If these scenarios are judged to be valid for past research (ostensibly supporting MFT), they must also be valid for our research—even though our research disconfirms MFT.

Scenario studies are not only the "most widely used by far" in the field (Graham et al., 2013, p. 70), but—critically—directly assess *intuitive* moral judgments regarding particular acts. Other MFT stimuli, especially the Moral Foundations Questionnaire (MFQ), do *not* assess intuitive moral judgments. Despite its name, the MFQ-Judgments scale assesses only the endorsement of general conservative values ("men and women have different roles to play in society," Graham et al., 2011). In other words, the MFQ-Judgments scale simply repackages political identification rather than assessing moral cognition—explaining why its effect on actual moral judgments is fully mediated by right wing authoritarianism (Kugler, Jost, & Noorbaloochi, 2014).

Even less useful for assessing intuitive moral judgment, the MFQ-Relevance scale asks participants to introspect about their own moral cognition ("what factors are relevant for your moral judgments?" Graham et al., 2011). Decades of research show that people lack introspective access to the reasons behind their judgments (Nisbett & Wilson, 1977), and as moral judgments are especially intuitive, deliberative moral reflection is unlikely to reveal anything more than lay theories and post hoc rationalizations (for a fuller explanation, see Haidt, 2001).

## Reliance on Participant Intuitions

MFT posits that harm and purity are distinct concerns, but the high correlation of these variables in participants' ratings (*r*s > .86) reveals a lack of distinctness. Graham criticizes this finding by suggesting that participants are unable to accurately identify harm and impurity. Why then do Graham and colleagues rely upon participants ratings of harm and impurity in their own studies? For example, the MFQ-Relevance instrument asks participants to rate whether "someone violated standards of purity and decency." If participant identifications of purity are judged to be valid for past research (ostensibly supporting MFT), they must also be valid for our research—even though our research disconfirms MFT.

Importantly, while the MFQ-Relevance scale asks people to reflect on the *reasons* behind their judgments (which we suggest previously is inappropriate), we merely asked people to rate the *presence* of purity or harm. This is no different from asking people to rate the presence of immorality (i.e., make a moral judgment). Moreover, in defining harm and purity for participants, we pulled words directly from the MFT "dictionaries" that ostensibly reflect laypeople's understandings of moral content (Graham, Haidt, & Nosek, 2009). If MFT "dictionary" words are judged to be valid for past research (ostensibly supporting MFT), they must also be valid for our research—even though our research disconfirms MFT.

Graham also criticizes our reliance on participant intuitions in constructing our new "purity" scenarios. We agree that our participant-generated cases (e.g., prostitution and stripping) differ substantially from MFT researcher-devised cases (adding a tail via plastic surgery; Haidt, 2012). However, we view the emphasis on everyday morality as a strength of our stimuli. As Graham notes, these naturalistic scenarios fail to independently activate harm and purity—but so too do the bizarre scenarios of MFT. We suggest that this reliable lack of independent activation stems not from specific scenarios but instead from a lack of moral modularity, consistent with domain-general dyadic morality.

## High Interconstruct Correlations

Not only did our data reveal extensive overlap between harm and purity, but ratings of harm and impurity were both highly correlated with judgments of immorality—that is, severity. Graham sees this as a problem, but it is problematic only for modular MFT. The harm-based template of dyadic morality suggests that perceived harm and perceived immorality *should* substantially overlap. That purity—as assessed with MFT items—is also correlated with harm and severity suggests that purity is either understood via harm (see Gray, Schein, & Ward, 2014) or is poorly defined, both of which challenge modular MFT.

## Domain-General Modularity?

The most surprising criticism leveled by Graham was that we incorrectly suggested that modular MFT was inconsistent with domain-general accounts. Graham asserts that MFT is "perfectly consistent with domain-general as well as domain-specific processes." This statement is a stark reversal for MFT. MFT researchers have repeatedly and explicitly argued *against* domain-general moral processes for more than a decade (Graham et al., 2013; Haidt & Joseph, 2004). Cognitive modules—encapsulated, domain-specific "switches"—are *by definition* opposed to domain general processes that cut across content (Cameron, Lindquist, & Gray, in press). How can Graham argue for the modularity of moral content—the specialness of purity per se—and accept that purity has no special effect on moral cognition beyond crosscutting domain-general dimensions?

Attempting to reconcile our domain-general findings with past published modular MFT claims, Graham suggests that there may be both "differences" and "similarities" across moral content. However, which exact differences and similarities MFT predicts are left vague. In order for a theory to be both falsifiable and useful (i.e., pragmatically valid, Graham et al., 2013), it must specify exactly and a priori when one pattern of results (e.g., differences)—versus its complete opposite (e.g., similarities)—are predicted to emerge. Unfortunately, Graham leaves these precise predictions unspecified, while past published formulations of MFT strongly argue only for differences (Graham et al., 2013).

### Pluralism and Parsimony

MFT presumes ownership of four claims: ‘‘nativism, culture, intuition, and pluralism’’ (Graham et al., 2013, p. 62). We challenge this presumed ownership. Dyadic morality *also* asserts that morality can be innate (nativism), learned (culture), and intuitive (Gray, Young, & Waytz, 2012). In contrast to the mischaracterization of Graham (2015), dyadic morality also embraces moral pluralism. Indeed, despite the anthropological roots of MFT, we suggest that dyadic morality is the true inheritor of pluralism because it also acknowledges *harm pluralism*—legitimate variations in perceived harm. Dyadic morality acknowledges that Brahmans legitimately see harm when burial rites are violated (Shweder, 2012), and U.S. conservatives legitimately see harm in homosexuality (Gray et al., 2014)—whereas harm-monist MFT rejects these perceptions as mere rationalizations (Haidt, 2001).

The pluralist dyad means that the ‘‘dyad versus MFT’’ debate is about the underpinnings of moral cognition—‘‘templates versus modules.’’ It is *not* about ‘‘parsimony versus pluralism,’’ as Graham suggests. Dyadic morality is both parsimonious *and* pluralist. In Shweder’s words, dyadic morality has ‘‘universality without uniformity,’’ by combining rich moral diversity with a common cognitive template. The variability of perceived harm gives this template flexibility, but it also yields a clear testable hypothesis: diverse moral judgments should reliably co-occur with intuitive perceptions of harm. If someone views something as immoral, they should also perceive it as harmful—a prediction supported by recent research (Gray et al., 2014).

## Conclusion

In sum, Graham’s comment provides a revised formulation of MFT that seems both internally inconsistent and impossible to reconcile with previous published accounts. Nevertheless, to the extent that MFT stands by its past-published scenarios, assumptions, and claims of modularity, we believe that we have disconfirmed MFT—and on its own terms.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

Cameron, C. D., Lindquist, K. A., & Gray, K. (in press). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*. Advance online publication. doi: 10.1177/1088868314566683

Graham, J. (2015). Explaining away differences in moral judgment: Comment on Gray & Keeney (2015). *Social Psychology and Personality Science*. Advance online publication. doi:10.1177/1948550615592242

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. Retrieved from http://doi.org/10.1016/B978-0-12-407236-7.00002-4

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. Retrieved from http://doi.org/10.1037/a0015141

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. http://doi.org/10.1037/a0021847

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenarios sampling bias raises questions about the foundation of morality. *Social Psychology and Personality Science*. Advance online publication. doi:10.1177/1948550615592241

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, *3*, 1–19. Retrieved from http://doi.org/10.1007/s13164-012-0112-5

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*, 1600–1615. Retrieved from http://doi.org/10.1037/a0036149

Gray, K., & Wegner, D. M. (2011). Dimensions of moral emotions. *Emotion Review*, *3*, 227–229.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124. Retrieved from http://doi.org/10.1080/1047840x.2012.651387

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon Books.

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

Kugler, M., Jost, J. T., & Noorbaloochi, S. (2014). Another look at moral foundations theory: Do authoritarianism and social dominance orientation explain liberal-conservative differences in ‘‘moral’’ intuitions? *Social Justice Research*, *27*, 413–431.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Shweder, R. A. (2012). Relativism and universalism. In D. Fassin (Ed.), *A companion to moral anthropology* (pp. 85–102). Hoboken, NJ: John Wiley & Sons, Ltd.

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

## Author Biographies

**Kurt Gray** is an assistant professor of psychology who studies mind perception and morality.

**Jonathan E. Keeney** is a PhD student who explores the interplay of moral cognition and real-world decision making.