Explaining Away Differences in Moral Judgment: Comment on Gray & Keeney (2015)

Jesse Graham

University of Southern California

in press, *Social Psychological and Personality Science*

Word count: 2632

Corresponding Author:
Jesse Graham
Department of Psychology
University of Southern California
3620 McClintock Ave.
Los Angeles, CA 90089
jesse.graham@usc.edu

Abstract

Moral judgments about harm vs. impurity differ in a number of non-superficial ways, as shown by dozens of studies, conducted by dozens of separate research labs, using a wide variety of methods and stimuli. Gray & Keeney (2015) attempt to explain away these differences by arguing that the "confounds" of severity and typicality may account for them all. This comment examines the evidence for this claim. Severity and typicality are undoubtedly important factors for moral judgment, but Gray & Keeney fail to demonstrate that they account for *any* (much less all) of the harm-impurity differences in the literature. Correlated ratings of "harm" and "impurity" are redundant with severity ($.93 \leq rs \leq .97$), merely tracking overall wrongness. The conclusion that harm and impurity judgments don't meaningfully differ at all, that all the functional and cognitive differences in the literature are "illusions" resulting from "confounds" and "sampling bias," is entirely unwarranted by the present studies.

Key words. Morality, Moral judgment, Monism, Pluralism, Evidence

Explaining Away Differences in Moral Judgment: Comment on Gray & Keeney (2015)

Is morality one thing or many? Do our moral reactions derive from a single process, or from multiple processes? Pluralist accounts of morality have proposed multiple different mechanisms of moral judgment, distinguished by different intuitive sensitivities (Moral Foundations Theory [MFT]; Graham, Haidt, Koleva, Motyl, Iyer, Wojcik, & Ditto, 2013), motivational foci (Model of Moral Motives; Janoff-Bulman & Carnes, 2013), relational contexts (Rai & Fiske, 2011), or value representations (Cushman, 2013). In contrast, monist accounts of morality propose that all moral judgments are essentially the same, deriving from a single process; for instance, dyadic morality (Gray & Schein, 2012) proposes that all moral judgments result from a process of matching events to a cognitive template consisting of harmful agent and suffering patient.

One piece of evidence for pluralist over monist accounts of moral judgment comes from experiments contrasting moral reactions to harm vs. impurity. These include the 53 studies cited by Gray and Keeney (2015), and many more besides: for example, judgments about harm vs. impurity (examined in isolation or in aggregates of harm/unfairness vs. betrayal/disrespect/impurity) respond in opposing ways to experimental manipulations (Napier & Luguri, 2013; Cornwell & Higgins, 2013), generate person- vs. situation-based attributions (Chakroff & Young, 2015), moderate domain-general effects such as omission/commission and means/byproduct biases (Descioli, Asao, & Kurzban, 2012) and, when experimentally primed, produce dramatic differences in important areas of thought and behavior (Day, Fiske, Downing, & Trail, 2014; Feinberg & Willer, 2013).

Gray and Keeney (2015, p. 2) argue that such differences between harm and impurity judgments "stem not from differences in moral content *per se*, but from biased sampling that confounds content with weirdness and severity." In this commentary I argue that this claim is not supported by the evidence. Measures of harm vs. impurity often differ on severity, weirdness, and countless other dimensions, but the authors fail to demonstrate that these dimensions fully account for any (much less all) of the dozens of harm/impurity differences found in the literature. I also discuss the theoretical leap from demonstrating the existence of such dimensions to claims that no empirical distinctions exist between harm and impurity, that such dimensions provide evidence against pluralist accounts of moral judgment, and that their existence supports monist accounts of a single moral process in which impurity is no more than "(perceived) harm involving sex."

I note at the outset that there is great value in identifying the many ways in which different kinds of moral judgments differ (using not just the constructs of MFT, but those of all models of morality), and identifying which of these differences contribute to the different behavioral and cognitive effects found when contrasting those constructs. This approach would move us toward a radically pluralist and unparsimonious account of morality, involving dozens of dimensions of difference and similarity between different types, contexts, and qualities of moral judgment, but I think such a move into the complexity and messiness of human morality is a worthwhile scientific endeavor.

Methodological problems

The authors claim that an entire subfield of moral psychology suffers from "confounds" and "biases," and that its findings may all be "illusions." These are serious accusations—what is the evidence for them?

Study 1. In Study 1 participants rated one of the four moral foundation measures used in Graham, Haidt, and Nosek (2009, Study 3) in terms of severity, weirdness, harm, and impurity. As expected, participants rated the harm items as more severe, and less weird, than the impurity items. It is completely uncontroversial that harm and impurity violations on average differ in terms of typicality and severity. MFT does not rest on a claim that all foundation-related judgments are equal in severity. And atypicality is in fact a primary feature of the Purity/degradation foundation; for instance, one of the Purity items in the Moral Foundations Questionnaire (MFQ; Graham et al., 2011) is to assess the moral relevance of "whether or not someone did something unnatural," as weird things can be seen as affronts to nature, or human nature, or God's plan (see also descriptions of the Purity foundation in previous MFT papers).

Moreover, typicality and severity aren't the *only* ways harm and impurity differ. One might also observe differences in grossness, funniness, vividness, contamination, self- vs. other-relevance, intentionality, speed of processing, or dozens of other dimensions. Again, there's value in such an approach, especially for figuring out why different kinds of moral judgments are processed differently. But simply finding differences on two dimensions and then concluding that there are (or may be) no other meaningful differences between harm and impurity goes far beyond the data.

Another critique of MFT is introduced in Study 1, by asking participants to rate how harmful and impure the violations are, and finding that these ratings correlated highly (r=.89). This is said to be the first test of the "surprisingly untested" claim that harm violations activate harm concerns and impurity violations activate impurity concerns. The reason for this surprising omission in the literature is that this is an extremely problematic method for testing the claim. You can't simply ask people to rate impurity or harm and conclude that they are able to self-

report to you which intuitive concerns have been activated. Such ratings of violations merely tap wrongness assessments; people will tell you something is bad in whatever way you allow them to tell you it's bad. Supporting this interpretation, Supplemental Tables 3 and 7 reveal nearly perfect correlations between ratings of severity and harm (r=.93, Study 1; r=.96, Study 2) and between ratings of severity and impurity (r=.96, Study 1; r=.97, Study 2). Thus these "harm" and "impurity" ratings are redundant with severity, merely tracking overall wrongness assessments. This is also why the more severe harm items are rated as more "impure" than the impurity items—again, these ratings of "impurity" are simply tracking overall severity and wrongness. Rather than a failed "manipulation check," it is a failed method for investigating which violations tap which concerns.

This critique is not a prioritizing of researcher intuitions over participant intuitions; it is a critique of the methods used to tap participant intuitions. Correlations between severity ratings using different adjectives tell us little about how distinct these moral concerns are, but correlations across people between moral judgments about harm and impurity violations can tell us more (e.g., how well can you predict people's impurity judgments from their harm judgments?). The harm-impurity correlation is r=.06 for the MFQ, r=.35 for the Moral Foundations Sacredness Scale (MFSS; Graham & Haidt, 2012), r=.23 for the MFSS adjusting for overall willingness to do things for money, r=.19 for the Moral Foundations Vignettes (MFV; Clifford et al., 2015), r=.06 for MFV items selected to match on severity and arousal (Dehghani et al., 2015), and r=.29 for the participant-generated naturalistic measures used by Gray and Keeney in Study 2. These correlations vary across formats, but are all substantially lower than the misleading ratings correlation that the authors conclude to be "casting doubt on the distinctness of these concerns" (p. 8).

Study 2. Study 2 finds the same severity/typicality differences, again using the same measure from Graham et al. (2009, Study 3). If the goal is to show that the field (or even MFT specifically) suffers from a "confound," why not use all four measures from this paper, or better yet why not use validated measures like the MFQ, MFSS, or MFV, or any of the other harm and impurity items researchers have developed? There does seem to be some selection bias here, as the items chosen to represent MFT are indeed the weirdest of the weird (I fully agree that tail-adding surgeries are highly atypical). But again, merely finding severity and typicality differences between harm and impurity violations is uncontroversial, and unproblematic for moral pluralism.

Study 2 then asks participants to generate their own examples of harm and impurity, and some of these examples are selected to form "naturalistic" measures of harm and impurity. The authors remark, "notably, no participant generated dog-eating, chicken-masturbation, urine-drinking, or soul-selling as purity violations" (p. 9). True, but participants also didn't generate dog-kicking, anthill-stomping, fat-shaming, or child-palm-sticking for harm, and they *did* generate incest, bestiality, masturbation, deviant sex, exhibitionism, urinating on someone, etc. for impurity. In general the participants generated items quite similar to those used in the literature.

However, some of the items selected for impurity involve violations of multiple foundations, such as rape (harm, impurity) and adultery (disloyalty, harm, impurity); although these items mitigate (but do not eliminate) differences in weirdness and severity, they also fail to distinguish harm content from impurity content. Violations triggering reactions built on multiple foundations (such as rape and adultery, which were generated for both harm and impurity) are in fact held up as evidence against "modular accounts," but this represents a misunderstanding of

such accounts; multiple MFT papers have discussed how particular values, virtues, and vices can be built on multiple intuitive foundations. Researchers creating MFT measures have sought to use items that relate to foundational concerns in isolation, just as someone creating a Big Five personality measure might want to avoid using items that tap both agreeableness and extraversion. Thus scenarios violating (or supporting) multiple foundations are not included in most MFT measures — if this is what is meant by "sampling bias," then it is a bias that affects every psychological measure intended to capture a particular construct with discriminant validity. Finally, as in Study 1, no attempt is made to demonstrate that severity or weirdness differences are responsible for any of the harm/impurity effects found in previous research.

Study 3. Study 3 employs an ideal design for testing the relative contributions of content, severity, and weirdness to one of the harm/impurity effects in the literature, manipulating each separately to examine differential effects on judgments of acts vs. character (Uhlmann & Zhu, 2013). Unfortunately, while the manipulation of severity seems reasonably valid, the weirdness manipulation is unconnected to the actual violations; more problematically, the manipulation of content fails to distinguish harm from impurity.

Adultery is chosen for the impurity violation, even though the authors themselves discuss (p. 10) how adultery violates multiple foundations (harm, loyalty, purity), and even though participants in Study 2 listed adultery as an example of both harm and impurity. To make things worse, the adultery scenario leaves ambiguous whether the adulterous behavior is mutually consensual adultery, or one-sided sexual assault ("Imagine a man French kisses and gropes someone who is not his wife after painting himself red and putting on a cape of human hair" does not exactly imply consent). Thus the manipulation of content does not contrast harm with impurity — it contrasts assault with sexual assault. The authors conclude that besides equating

scenarios on severity and weirdness, future studies should "ensure that impurity scenarios activate impurity concerns, but *not* harm concerns, and vice versa for harm scenarios" (p. 20). This is exactly what a study contrasting assault with sexual assault fails to do.

Although Study 3 does not fairly contrast harm vs. impurity, even here the authors are forced to conclude "as in Study 2, this suggests that severity and weirdness likely do not account for all differences between harm and impurity scenarios" (Gray & Keeney, p. 17). This admission, which is in stark contrast to the claims made throughout the rest of the paper, is supported by a recent study of a related harm/impurity effect. Investigating differential activation of situation- vs. person-based attributions, Chakroff and Young (2015, Study 2) included ratings of severity (perceived wrongness) and weirdness (perceived abnormality) as potential mediators of the harm/impurity effect. Both of these factors partially mediated the effect, but the direct effect of content domain remained when they were included.

Theoretical interpretation problems

Despite not providing any evidence that severity or weirdness explain any of the harm/impurity effects in the literature, Gray and Keeney conclude that these dimensions 1) represent "confounds" produced by "sampling bias," 2) raise "doubts" and "questions" about pluralist accounts of morality, and 3) support a monist harm-based account. First, the uses of "confounds" and "sampling bias" are not justified by the data. These are serious charges (that theories are wrong, previous effects of a whole subfield are spurious, etc.); absent evidence to back them up, they are primarily arguments from innuendo. Conceptually, I think harm and impurity violations really do vary by severity and typicality (especially within secular Western cultures), so it's not clear why these differences are confounds — they are two of the many ways these two kinds of moral violations actually differ.

Violations that are impure are said to be "just weird," nothing other than atypical (and less severe) harm violations. Or, alternately, "perhaps we can simply define impurity as '(perceived) harm involving sex'" (Gray & Keeney, 2015, p. 19) — this definition is particularly surprising, given that in the most widely-used MFT measure, the MFQ, only one of the six Purity items contains sexual content ("Chastity is an important and valuable virtue"), and even in the MFT measure used in the current studies only one of the five impurity items contains sexual content. Both of these reconceptualizations represent attempts to shoehorn impurity into a monist harm-based account, wherein all moral judgments follow the same agent-harming-patient template matching process. This is not to say that harm violations cannot be the most severe, or even the most prototypical, moral violations. But there is a big conceptual leap from identifying harm as more severe and typical than impurity to suggesting that harm is the only true moral concern, and that impurity (and lying, cheating, betrayal, etc.) must be cast as special cases of this one moral concern (a la "harm involving sex"). Throughout Gray and Keeney's argument, domain-general dimensions are equated with this monist account, as if such dimensions contradict with pluralist (or "modular") accounts. But moral pluralism (and MFT specifically) is perfectly consistent with domain-general as well as domain-specific processes. Similarly, constructionist approaches to morality can be quite valuable (Cameron et al., in press), but these too are compatible with pluralist accounts, far from the exclusive purview of moral monism.

Conclusion

Moral judgments about harm vs. impurity differ in a number of non-superficial, psychologically important ways, as shown by dozens of studies, conducted by dozens of separate research labs, using a wide variety of methods and stimuli. Gray and Keeney attempt to explain away these differences by arguing that the "confounds" of severity and typicality may account

for all of them. Severity and typicality are undoubtedly important factors for moral judgment, but the authors fail to demonstrate that they account for *any* of the harm/impurity differences in the literature. And figuring out why moral reactions to harm and impurity differ is certainly a worthwhile goal, but to conclude that harm and impurity judgments don't meaningfully differ at all, that all the functional and cognitive differences in the literature are "illusions" resulting from "confounds" and "sampling bias," is completely unwarranted by the present studies. The authors argue from a mutually exclusive dichotomy between domain-specific and domain-general accounts of morality, as though there can only be similarities OR differences between different kinds of moral judgments. But of course there can be (and based on the existing evidence, are) both similarities and differences. Evidence for cognitive differences does not preclude there also being similarities, and evidence for cognitive similarities does not preclude there also being differences.

Nevertheless, in closing I want to acknowledge that these theoretical disagreements are healthy and important for moral psychology. It is a good thing that some scientists are focused on finding differences, and others are focused on finding similarities, across types and contexts of moral judgment. And while I think that a pluralist approach allowing for both domain-general and domain-specific factors can best account for and explain the complexity of moral life, there is also value in a plurality of theoretical and methodological approaches in pursuit of this goal.

Acknowledgement

Thanks to Alek Chakroff, Jon Haidt, Dan Molden, Brian Nosek, Josh Rottman, Piercarlo Valdesolo, and Liane Young for helpful feedback on a draft 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*.

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*, 1–21.

Cornwell, J. F., & Higgins, E. T. (2013). Morality and its relation to political ideology: The role of promotion and prevention concerns. *Personality and Social Psychology Bulletin, 39*, 1164-1172.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*, 273-292.

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.

Dehghani, M., Johnson, K. M., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2015). Purity homophily in social networks. Manuscript submitted for publication.

DeScioli, P., Asao, K., & Kurzban, R. (2012). Omissions and byproducts across moral domains. *PLoS ONE, 7*, e46963.

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science, 24,* 56-62.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. H. (2013). Moral Foundations Theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47,* 55-130.

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.

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* (pp. 11-31). New York: APA Books.

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.

Gray, K. & Keeney, J. (2015). Impure, or just weird? Scenario sampling bias raises questions about the foundations of morality. *Social Psychology and Personality Science.*

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.

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.

Napier, J. L., & Luguri, J. B. (2013). Moral mind-sets: Abstract thinking increases a preference for "individualizing" over "binding" moral foundations. *Social Psychological and Personality Science, 4,* 754-759.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118,* 57–75.

Uhlmann, E. L., & Zhu, L. (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science*, *5*, 279-285.