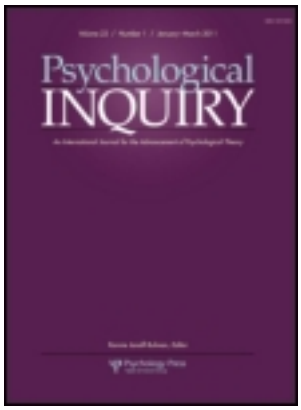

Psychological Inquiry: An International Journal for the Advancement of Psychological Theory

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hpli20

The Unbearable Vagueness of "Essence": Forty-Four Clarification Questions for Gray, Young, and Waytz

Jesse Graham [a] & Ravi Iyer [a]

[a] Department of Psychology, University of Southern California, Los Angeles, California

Available online: 31 May 2012

To cite this article: Jesse Graham & Ravi Iyer (2012): The Unbearable Vagueness of "Essence": Forty-Four Clarification Questions for Gray, Young, and Waytz, Psychological Inquiry: An International Journal for the Advancement of Psychological Theory, 23:2, 162-165

To link to this article: http://dx.doi.org/10.1080/1047840X.2012.667767

Psychological Inquiry, 23: 162–165, 2012

ISSN: 1047-840X print / 1532-7965 online
DOI: 10.1080/1047840X.2012.667767

Psychology Press
Taylor & Francis Group

The Unbearable Vagueness of "Essence": Forty-Four Clarification Questions for Gray, Young, and Waytz

Jesse Graham and Ravi Iyer
Department of Psychology, University of Southern California, Los Angeles, California

To make the argument that all morality is essentially one thing, Gray, Young, and Waytz employ a series of helpful analogies, portraying morality as a bull, an elephant, a dog, a Necker cube, H_2O, a university, an invisible triangle, and the Grand Canyon Skywalk. This impressive metaphoric diversity illustrates just how difficult it is to fit something as rich and complex as human morality into a single characterization. It also illustrates the authors' vagueness about what exactly is being argued by "essence."

The target article makes three claims. The first claim, "perceptions of mind are linked to moral judgments," is well supported by a comprehensive overview, and a strong case is made that mind perception and morality are "closely linked," "connected," and even "naturally connected." This claim is uncontroversial—it's difficult to imagine any theoretical take on morality that doesn't see mind perception (and social cognition more generally) as closely interwoven with it, especially if mind perception includes "perceptions of group mind" (p. 114). For instance, the idea that moral thinking is for social doing (Haidt, 2007) implies that perceptions of the minds of others will be a crucial aspect of morality. The second claim, "dyadic morality uniquely accounts for the phenomena of dyadic completion . . . and moral typecasting," is also well supported: These two phenomena represent the birthplace of the dyadic morality theory, and no other theory accounts for them so well. However, the hyperbolic third claim, "all moral transgressions are fundamentally understood as agency plus experienced suffering," is not well supported (by data or argument); more problematically, it never becomes clear what exact claim is being made by terms like "essence," "fundamentally understood," "understood through the lens of," and so on. In this commentary we ask Gray, Young, and Waytz an overarching question—What does "essence" mean?—as well as four more specific sets of questions that cannot be answered until the meaning of "essence" is clear.

Questions 1 to 9: What Does It Mean to Say Mind Perception Is the Essence of Morality?

Gray, Young, and Waytz (this issue) do an admirable job of laying out the evidence for a strong link between mind perception and morality. But they make explicit that they want to claim more than just a strong link: "Many researchers have shown that mental state attribution is important to morality, but here we explore whether mind perception is the *essence* of morality" (p. 103). This certainly seems like a bolder claim, but what exactly is the step from importance to essence? What new claim is being introduced? Is the claim best characterized by the metaphor of Picasso's bull (here's one of many elegant ways of picturing the most important and prototypical features of morality, with as few strokes as possible) or is it closer to the metaphor of H_2O (concerns such as social justice and moral disgust might seem different on the surface, but in actual, testable reality they are the exact same thing)?

Is the claim that mind perception is a *necessary precursor* to all moral judgments? This is a potentially useful and testable claim, but the authors provide no specific evidence for it (e.g., demonstrations that perceptions of mind precede affect), and it's not clear that this stronger claim is being made. For instance, when discussing the role of affect in moral judgments, they make the softer claim "*Most often*, [affect] seems to be triggered by perceiving a mind" [emphasis added] (p. 115) and conclude with the even softer claim that both cognitive and affective components are simply *linked* to mind perception. By saying that mind perception is the *essence* of morality, are the authors claiming that it is not only very important for morality, but that it is the *most* important factor? Is mind perception always (for all people, in all contexts, regarding all content areas) more central and important to morality than all other factors, like affect, consequences, culture, or rule violations? At points in the target article it seems the claim is that mind perception is the *best* lens through which to understand all moral judgments. It's unclear how one could test this theoretical superiority claim, but perhaps this type of claim is more appropriately advanced by argument than data. However, despite the impressive evidence offered for the links between mind perception and morality, we remain unconvinced that mind perception is always the best way to understand moral judgments or concerns. Take, for instance, the robust finding that moral judgments become harsher in the presence of incidental disgust (e.g., dirty desk, chewed

pens, or flatulent aromas; Schnall, Haidt, Clore, & Jordan, 2008). Does mind perception *best* illuminate why this takes place? Does it offer the *best* way to understand this phenomenon? If this is not the intended meaning of "essence," and nor is the claim that mind perception is a necessary precursor to moral judgments, then we are not sure what the essence claim does beyond assert that "mind perception is very important for morality." Most if not all of these open questions would be resolved if the authors specified what exact claims are being made by the use of "essence" in their article's title.

Questions 10 to 23: What Does It Mean to Say Dyadic Harm Is the Essence of Morality?

Despite the target article's title, most of its space is devoted to arguing for the essentiality of a specific "template" of mind perception—the dyadic combination of intentionally harmful agent and suffering patient—rather than mind perception in general. This introduces more uncertainty about what is being claimed: Is the essence of morality perception of minds, or perception of these two specific kinds of minds? Does this give us a more specific necessary precondition for all moral judgments, and if so, do all four components (dyad, intention, harm, suffering) need to be perceived for a moral judgment to occur? What if some of these features are perceived but others are not, as in the perception of dyadic dishonesty (intentionally deceitful agent, no harm done, unsuffering deceived patient)? If the lie is harmless ("I love that sweater"), then most people would probably judge it less harshly than a harmful lie. But the claim here isn't just that harm perceptions (among others) are important to moral judgments, it's that they alone are *essential* for moral judgments: "A dyadic template suggests not only that perceived suffering is tied to immorality, but that all morality is understood through the lens of harm" (p. 108). Again, is this a process claim of necessary precondition, whereby all moral judgments must be reached via the perception of dyadic harm? Or is the claim a softer one in which "essence" only means "template," and so (most) moral judgments (roughly) fit the prototype of dyadic harm (usually)?

These questions can also be answered by clarifying exactly what "essence" means. But vagueness will remain until the word "harm" is defined as well. At times the authors seem to mean something concrete like "intentionally caused physical or emotional suffering," but other times "harm" is stretched to be nearly synonymous with anything morally bad. For instance, in the case of perceived dyadic dishonesty, the authors could claim that even though no suffering was directly experienced by the unknowingly deceived patient, the social contract was harmed by the lack of honesty, or the deception *can* lead to harm at some point in the future. This is the rhetorical strategy employed in the section titled "Concerns About Suffering Underlie Different Domains"—a bold, specific, and potentially useful empirical claim. However, this section just asserts that nonharm violations *can* lead to suffering ("can" is used 10 times in this 10-sentence paragraph). What is the claim that these "can" arguments support? One can come up with ways that harms *can* result from harmless violations, but does this mean that suffering concerns *must* "underlie" reactions to these violations? For instance, purity violations can lead to suffering, as when promiscuous sex results in a burning sensation. Does this mean that when people make moral judgments about promiscuity they do so only by reference to the physical or emotional suffering that can result from it, and the intentionally harmful agents who engage in such behavior? Do people really only judge promiscuity as wrong because of its similarity to the prototype of dyadic harm? Another example: Fairness violations can lead to suffering, when an unfairly distributed resource is needed. What if the resource is not needed, such as one child getting a surprise present when her sister gets nothing? If the sister feels that this is wrong, does she do so only by reference to a perception of herself as a suffering patient and the gift-giver as an intentionally harmful agent, or does she simply perceive unfairness? Again, is the claim here a process model, in which anything leading to a moral judgment must proceed via reference to dyadic harm? Is the claim that there is only one moral judgment, only one moral intuition—an intuitive response to dyadic harm—and some harmless things just happen to trigger it?

As with mind perception in general, the authors make a strong case that intentional harm and perceived suffering are both very important for moral judgments. But this is presented as evidence that dyadic harm is the essence of moral judgment: "If the essence of morality is captured by the combination of harmful intent and painful experience, then acts committed by agents with greater intent and that result in more suffering should be judged as more immoral" (p. 106). This statement may be true, but that doesn't mean that its reverse is also true. That is, the fact that greater intent and suffering lead to greater perceived immorality does not provide evidence that these are the essence of morality, any more than fart sprays increasing severity of moral judgments provide evidence that flatulence is the essence of morality.

Questions 24 to 33: How Is the Theory Falsifiable?

To know how the theory of dyadic harm morality can be falsified, we need to clarify what specific claims are being made. This problem is by no means unique to this theory – it is also not fully clear how most

other theories of morality (e.g., moral foundations, moral components, universal moral grammar) can be falsified. Assuming the authors are making the bolder claims they occasionally step back from, and assuming we're right in our interpretations that these bolder claims involve a kind of process model and necessary precursor argument, then we have some suggestions about how dyadic harm morality can be falsified. The theory seems to us to imply that cases like Wheatley and Haidt's (2005) hypnotic suggestion of "disgust," and Schnall et al.'s (2008) manipulations of incidental disgust in the environment, could only be increasing moral judgments via a process of increased perceptions of intentional harms and suffering patients. This is an empirically testable claim, and could perhaps be tested using habituation procedures to deactivate concepts related to harm and suffering—if disgust is only affecting moral judgments via these concepts, then this should nullify the disgust effects. Similarly, depending on what is known about the time-course of mind perception, the claims that it is a *necessary* precursor to moral judgment could be falsified by showing signs or consequences of moral judgment before signs or consequences of mind perception were apparent. Such demonstrations would be difficult to achieve, if they are possible at all. But the more pressing question is, What would the authors consider to be disconfirming evidence of the claims of their theory? This of course cannot be answered until we know what these claims are.

Related to the question of falsifiability. What can and should count as evidence for the theory? The authors offer plenty of convincing evidence for the links and importance of mind perception, intention, harm, and suffering to morality—but what would provide evidence for their claims of essence beyond mere links or importance? Without knowing what "essence" means, how can we test whatever claims this step adds?

Gray, Young, and Waytz deserve genuine credit for attempting to apply the dyadic harm model in realms other than the phenomena it was created to explain. Many other parsimonious, elegant accounts of human morality are born in a particular set of phenomena (responses to trolley dilemmas, behavior in economic games), and then those same phenomena are offered as evidence for the theories. So we find it refreshing that the authors apply the theory in areas where it seems least likely to work (e.g., moral disgust, honor killings, character judgments). However, these attempts fail in the execution: Precisely when it seems like the authors are going to show how dyadic harm *explains* morality at the different "levels" of community, character, and components, they give up on arguing for harm or dyads at all and instead fall back on the weaker, noncontroversial argument that mind perception is *linked* to all these instances of morality. It's not very surprising that mind perception *occurs* when assessing someone's character, or when focused on the group as locus of moral concern, but where is the evidence or even the argument that dyadic harm unites and explains these as their essence? What about the harsh moral character judgments in cases of less harmful (Tannenbaum, Uhlmann, & Diermeier, 2011) or even harmless (Inbar, Pizarro, & Cushman, 2012) violations? What about positive moral judgments? Do all responses to prosocial action and character virtues rely on reference to dyadic helping? What about the reduced role of intent in moral judgments about harmless Purity violations (Young & Saxe, 2011)? For that matter, what about all of the theory-contradicting moral disgust evidence the authors cite on page 110? The authors dismiss this body of evidence by pointing out that "disgust initially evolved to protect people from bodily harm . . . and so the experience of moral disgust can be seen as a heuristic for potential suffering" (p. 110). By this logic, *anything* that conferred a survival advantage is a "heuristic for potential suffering," because it reduced the pain of not leaving offspring and/or dying, and death after all is really quite harmful. To show that dyadic harm is the essence of all morality, the authors need to do more than show how mind perception is *linked* to these "exceptions," or how harms *can* arise from harmless violations. But again, it all depends on what "essence" means.

Questions 34 to 44: What Is the Pragmatic Validity of the Theory?

As we previously noted, there are many parsimonious, elegant characterizations of the moral sense, and it is striking how different these can be in both theoretical approach and empirical focus. How to adjudicate among them? Is morality best characterized as dyadic harm, or as fairness (Baumard, André, & Sperber, in press), or as universal grammar (Mikhail, 2007), or as something else? Which of these pictures is most elegant? This does not strike us as an empirical question. But what if we asked which picture is most useful? Here there may be enough traction to be able to eventually provide an empirical answer.

In the context of validating a pluralistic measure of moral foundations (Graham et al., 2011), we employed the concept of pragmatic validity in homage to William James:

> Pragmatism asks its usual question. "Grant an idea or belief to be true," it says, "what concrete difference will its being true make in anyone's actual life? How will the truth be realized? What experiences will be different from those which would obtain if the belief were false? What, in short, is the truth's cash-value in experiential terms?" (James, 1907/1998, p. 97)

What is the "cash-value" of the dyadic harm theory of morality? What new understanding of human morality does it bring to science, and what new questions does it allow researchers to ask? The authors explicitly address these questions in the section titled "Novel Phenomena of Dyadic Morality," and here we are entirely convinced of the theory's usefulness in conceptualizing and explaining dyadic completion and moral typecasting. Explorations of the interplay between moral agency and patiency have been a concrete and useful addition to moral psychology. Most of the target article, though, is devoted to arguments that mind perception and dyadic harm are not just important but essential for understanding morality. The "unification" of moral psychology is offered as a primary benefit of the dyadic-harm approach, and yet the section on unification of levels of analysis only provides arguments that mind perception in general (not dyadic harm specifically) is linked to the different levels. What new understandings or insights does it provide about morality to say that it always involves some perception of other minds? What new hypotheses are made possible by the claim of "essence" as opposed to mere importance? The authors conclude that the theory "accounts for diverse findings in moral psychology" (p. 118), but this was never shown. Dyadic morality has a very useful account of moral typecasting and dyadic completion, the phenomena out of which the theory was born. But it's not yet clear how it accounts for other moral phenomena in a similarly useful way.

The more bold and novel claim, that all moral judgment necessarily involves some reference to dyadic harm, could be quite useful in other respects. For instance, it would provide a clear and (potentially, at least) measurable criterion for calling a judgment, value, or attitude moral as opposed to nonmoral. If there is a dispute about whether, say, purity concerns are really moral or not, this can now be an empirical question, not just semantic. Although we think there is scant evidence for this stricter claim, we see it as a potentially useful, as-yet-unsupported hypothesis, and one of the most promising future avenues for this approach to studying morality.

Conclusion

Although our commentary has concentrated on areas of potential disagreement, we raise these 44 questions not simply to critique but also as genuine clarification questions. We fully expect that the authors will have good answers to these questions (and we won't perceive any harmful intent if they only have space in their reply to answer 41 or 42 of them). Despite our skepticism about many of the theory's claims, we think it is worthwhile to boldly push a theory as far as it will go, and that something valuable is learned in that process. The need to clarify claims applies to many theories of morality (e.g., what exactly is claimed by "foundation," or by "grammar," etc.), and increasing specification will help determine where the different theories make different predictions, and how to empirically test them. By clarifying what exact claims are being made by the use of "essence" in the target article, we can begin the work of determining whether those claims are correct and what additional pragmatic value they can bring to moral psychology.

Note

Address correspondence to Jesse Graham, Department of Psychology, University of Southern California, 3620 South McClintock Avenue, SGM 501, Los Angeles, CA 90089-1061. E-mail: jesse.graham@usc.edu

References

Baumard, N., André, J. B., & Sperber, D. (in press). A mutualistic approach to morality. *Behavioral and Brain Sciences.*

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.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998–1002.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*, 52–62.

James, W. (1998). *Pragmatism: A new name for some old ways of thinking and The Meaning of Truth.* Cambridge, MA: Harvard University Press. (Original work published 1907)

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences, 11*, 143–152.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096–1109.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47*, 1249–1254.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*, 780–784.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*, 202–214.